

Refinement of the Sphere–Packing Bound: Asymmetric Channels

Yücel Altuğ *Student Member* and Aaron B. Wagner *Member*

Abstract

We provide a refinement of the sphere-packing bound for constant composition codes over asymmetric discrete memoryless channels that improves the pre-factor in front of the exponential term. The order of our pre-factor is $\Omega(N^{-\frac{1}{2}(1+\epsilon+\rho_R^*)})$ for any $\epsilon > 0$, where ρ_R^* is the maximum absolute-value subdifferential of the sphere-packing exponent at rate R and N is the blocklength.

I. INTRODUCTION

Characterizing the interplay between the rate, blocklength and error probability of the best block code(s) on a discrete memoryless channel (DMC) is a central problem of information theory. Although it has been investigated since the early days of the field [1]–[12], it is still an active research topic [13]–[23]. In a broad sense, there are two approaches to this problem:

- (i) *Finite blocklength results*: Because of the significance of the short to moderate blocklengths in practice, one can seek finite blocklength bounds on the error probability for a given rate. This can be done for a general class of channels (e.g. [13], [14]) or particular channels (e.g. [16, Theorem 35], [16, Theorem 38]). Although these bounds are useful to assess the performance of practical codes, they are typically not conceptually illuminating.
- (ii) *Asymptotic results*: An alternative to finite blocklength results is resorting to an infinite blocklength limit to derive more insightful results. Although such results do not give “hard” bounds that are valid for small blocklengths, they do provide memorable rules of thumb. Furthermore, finite blocklength bounds can often be extracted from their proofs.

We shall adopt the asymptotic approach in this paper. There exist several asymptotic regimes in the literature, such as *error exponents* (e.g. [5]–[8]), the *normal approximation* (e.g. [4], [16]) and *moderate deviations* (e.g. [17], [18]). We call error exponents, the normal approximation and moderate deviations the *small error probability*, *large error probability*, and *medium error probability* regimes, respectively. In this paper, our focus will be on the small error probability regime. This regime not only has theoretical significance, but has practical value in those applications, such as data storage, that require extremely small error probabilities without the aid of feedback.

Classical asymptotic results on the small error probability regime focus only on determining the exponents. In particular, until recently, the tightest pre-factor for the upper bound on the error probability was $\Theta(1)$, due to Fano [5] and Gallager [6]. The best pre-factor in the lower bound for constant composition codes was $\Theta(N^{-|\mathcal{X}||\mathcal{Y}|})$, due to

The material in this paper was presented in part at the 2012 IEEE International Symposium on Information Theory (ISIT), Boston, MA. The authors are with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY, 14853, USA. E-mail: ya68@cornell.edu, wagner@ece.cornell.edu.

Haroutunian [8], [12, Theorem 2.5.3], where $|\mathcal{X}|$ and $|\mathcal{Y}|$ are the cardinalities of the input and output alphabets, respectively. (The original sphere-packing bound, derived by Shannon-Gallager-Berlekamp [7, Theorem 2], had an $\Theta(e^{-\sqrt{N}})$ pre-factor.) Clearly, there is a considerable gap between the orders of the pre-factors in the upper and lower bounds.

Recently, the authors have been working to reduce the gap between the pre-factors. The recent paper [21] considers symmetric channels and refines the sphere-packing lower bound by proving a pre-factor of $\Theta(N^{-\frac{1}{2}(1+|E'_{\text{sp}}(R)|)})$, where $E'_{\text{sp}}(R)$ is the slope of the sphere-packing exponent at point R . The paper [24] proves a refined random coding bound with a pre-factor of $O(N^{-\frac{1}{2}(1+\epsilon+\tilde{\rho}_R^*)})$ for any $\epsilon > 0$, for a broad class of channels, which includes all positive channels with positive dispersion. Here, $\tilde{\rho}_R^*$ is related to the subgradient of the random coding exponent, which reduces to $|E'_r(R)|$ for the case of completely symmetric or positive and symmetric channels; hence the optimal order of the pre-factor is determined, up to the sub-polynomial terms.

This work is a generalization of [21] to asymmetric channels. We prove a lower bound for constant composition codes with a pre-factor of $\Omega(N^{-\frac{1}{2}(1+\epsilon+\rho_R^*)})$ for any $\epsilon > 0$, where ρ_R^* is the maximum absolute-value subgradient of the sphere-packing exponent. While the essential approach is similar to that of [21], the asymmetry of the channel results in a significantly more involved argument compared to its symmetric counterpart. Although some improved finite- N bounds could be extracted from the proofs in this paper, the task of optimizing these bounds and numerically comparing them to the existing bounds is not pursued, since we focus on the asymptotic characterization.

An analogy to sums of i.i.d. random variables is instructive. The small, medium, and large error probability regimes of channel coding correspond to large deviations, moderate deviations, and central limit theory of i.i.d. sums of random variables, respectively. Along the same analogy, the setup of this work resembles the *exact asymptotics* problem in large deviations [25], [26, Theorem 3.7.4]. This problem aims to determine the pre-factor of the exponentially vanishing term in the large deviations theorem. Bahadur and Ranga Rao [25] characterized this pre-factor, $\Theta(1/\sqrt{N})$, including the constant, under some regularity conditions. Their result, in the form stated by Dembo and Zeitouni [26, Theorem 3.7.4], is the following:

Theorem 1.1: (Bahadur-Ranga Rao) Let μ_N denote the law of $\hat{S}_N = \frac{1}{N} \sum_{i=1}^N Z_i$, where Z_i are i.i.d. real valued random variables with logarithmic moment generating function $\Lambda(\lambda) = \log \mathbb{E}[e^{\lambda Z_1}]$. Consider the set $A = [a, \infty)$, where $a = \Lambda'(\eta)$ for some positive $\eta \in \{\lambda : \Lambda(\lambda) < \infty\}^\circ$. If the law of X_1 is non-lattice, then $\lim_{N \rightarrow \infty} J_N \mu_N(A) = 1$, where

$$J_N := e^{N\Lambda^*(a)} \eta \sqrt{\Lambda''(\eta) 2\pi N}$$

and $\Lambda^*(\cdot)$ is the Fenchel-Legendre transform of $\Lambda(\cdot)$.

If X_1 is a lattice random variable, then the order of the pre-factor is the same, but the constant is different. Hence, $\Theta(N^{-\frac{1}{2}})$ is the correct order of the pre-factor for i.i.d. sums of random variables, and this factor will appear in our channel coding result. When one reduces the error event of a code to a sum of independent random variables, however, the threshold a must vary slightly with N , as will be evident in the sequel. This complicates the proof

by preventing one from directly applying the Bahadur-Ranga Rao result (the random variables are also not i.i.d. but merely independent). More importantly, the slow variation of the threshold changes the order of the pre-factor slightly, to include the slope term mentioned above.

The remainder of the paper is devoted to the statement and then the proof of our result.

II. NOTATION, DEFINITIONS AND MAIN RESULT

A. Notation

Boldface letters denote vectors, regular letters with subscripts denote individual elements of vectors. Furthermore, capital letters represent random variables and lowercase letters denote individual realizations of the corresponding random variable. Throughout the paper, all logarithms are base- e unless otherwise is stated. For a finite set \mathcal{X} , $\mathcal{P}(\mathcal{X})$ denotes the set of all probability measures on \mathcal{X} . Similarly, for two finite sets \mathcal{X} and \mathcal{Y} , $\mathcal{P}(\mathcal{Y}|\mathcal{X})$ denotes the set of all stochastic matrices from \mathcal{X} to \mathcal{Y} . Given any $P \in \mathcal{P}(\mathcal{X})$, $\mathcal{S}(P) := \{x \in \mathcal{X} : P(x) > 0\}$. $\mathbb{1}_{\{ \cdot \}}$ denotes the standard indicator function. Given two probability measures λ_1, λ_2 , $\lambda_1 \ll \lambda_2$ means ‘ λ_1 is absolutely continuous with respect to λ_2 ’ and $\lambda_1 \equiv \lambda_2$ is equivalent to saying $\lambda_1 \ll \lambda_2$ and $\lambda_2 \ll \lambda_1$. Φ (resp. ϕ) denotes the distribution (resp. density) of the standard Gaussian random variable. For a set S ; S^c , $\text{cl}(S)$, S° and $\text{ri}(S)$ denotes complementary set, closure, interior and relative interior, respectively. $\mathbb{R}_+, \mathbb{R}^+$ and \mathbb{Z}^+ denotes the set of non-negative real numbers, positive real numbers and positive integers, respectively.

B. Definitions

Throughout the paper, let W be a DMC satisfying¹ $R_\infty < C$. For any $P \in \mathcal{P}(\mathcal{X})$, define

$$E_{\text{SP}}(R, P) := \min_{V \in \mathcal{P}(\mathcal{Y}|\mathcal{X}) : I(P;V) \leq R} D(V||W|P),$$

and $E_{\text{SP}}(R) := \max_{P \in \mathcal{P}(\mathcal{X})} E_{\text{SP}}(R, P)$.

The following can be shown² to be the *maximum absolute value subgradient* of the sphere packing exponent at point R

$$\rho_R^* := \max_{P \in \mathcal{P}(\mathcal{X}) : E_{\text{SP}}(R, P) = E_{\text{SP}}(R)} |E'_{\text{SP}}(R, P)|, \quad (1)$$

where $E'_{\text{SP}}(R, P)$ denotes the slope³ of $E_{\text{SP}}(\cdot, P)$ at point R .

Given any (N, R) code (f, φ) , let $e(f, \varphi)$ (resp. $e_m(f, \varphi)$) denotes its maximal error probability (resp. error probability of the m -th message).

¹For the definition of R_∞ , see [12, pg. 170].

²Since $E_{\text{SP}}(\cdot, P)$ is convex for all $P \in \mathcal{P}(\mathcal{X})$, $E_{\text{SP}}(\cdot, \cdot)$ is continuous on $(R_\infty, \infty) \times \mathcal{P}(\mathcal{X})$ (cf. Lemma F.2 in the Appendix F) and $\mathcal{P}(\mathcal{X})$ is compact, one can invoke the characterization of the subdifferential of the maximum function (e.g. [27, Theorem 2.87]) to deduce that $\partial E_{\text{SP}}(R) = \text{conv}(\cup_{P: E_{\text{SP}}(R, P) = E_{\text{SP}}(R)} \{\partial E_{\text{SP}}(\cdot, P)(R)\})$, where $\text{conv}(\cdot)$, $\partial E_{\text{SP}}(R)$ and $\partial E_{\text{SP}}(\cdot, P)(R)$ denotes the *convex hull*, *subdifferential of $E_{\text{SP}}(\cdot)$ at point R* and *subdifferential of $E_{\text{SP}}(\cdot, P)$ at point R* , respectively. This observation, coupled with the differentiability of $E_{\text{SP}}(\cdot, P)$, i.e. Proposition 3.3, and the continuity of $E'_{\text{SP}}(R, \cdot)$, i.e. Proposition 3.4, suffices to conclude the claim.

³One can show that $E_{\text{SP}}(R, P)$ is differentiable with respect to R , for given P , provided that $R_\infty < R < C$ and $E_{\text{SP}}(R, P) > 0$, e.g. Proposition 3.3.

Let \mathcal{Z} be a finite set and $Q, \hat{Q} \in \mathcal{P}(\mathcal{Z})$. A *deterministic hypothesis test*, $T : \mathcal{Z} \rightarrow \{0, 1\}$, over the set \mathcal{Z} in which Q is the null hypothesis (H_0) and \hat{Q} is the alternate hypothesis (H_1) is defined as

$$T(z) = \begin{cases} 0, & \text{if } z \in \mathcal{U}_T, \\ 1, & \text{if } z \in \mathcal{U}_T^c, \end{cases}$$

where $\{\mathcal{U}_T, \mathcal{U}_T^c\}$ are called the *decision regions* of the test. Let $\mathcal{T}(Q, \hat{Q})$ denote the set of all deterministic tests between Q and \hat{Q} . The error probabilities associated with T are defined as $\alpha_T := Q\{\mathcal{U}_T^c\}$ and $\beta_T := \hat{Q}\{\mathcal{U}_T\}$. For any $r > 0$, define

$$\alpha_{Q, \hat{Q}}^*(r) := \min_{T \in \mathcal{T}(Q, \hat{Q}) : \beta_T \leq e^{-r}} \alpha_T. \quad (2)$$

C. Main Result

Theorem 2.1: Consider any $R \in (R_\infty, C)$ and $\zeta \in \mathbb{R}^+$. Then, for any sufficiently large N , depending on R , W and ζ and any (N, R) constant composition code (f, φ) ,

$$e(f, \varphi) \geq K \frac{e^{-NE_{\text{sp}}(R)}}{N^{\frac{1}{2}(1+(1+\zeta)\rho_R^*)}}, \quad (3)$$

where $K \in \mathbb{R}^+$ is a constant that depends on R , W and ζ .

III. PROOF OF THEOREM 2.1

A. Overview

There are at least three proofs of the sphere-packing bound in the literature: that of Shannon-Gallager-Berlekamp [7], Haroutunian [8] and Blahut [28]. Of these, Blahut's argument seems to be the most natural starting point for obtaining improved pre-factors, as it allows one to convert the error event of a code into an event involving a sum of i.i.d. random variables, to which one can apply the Bahadur-Ranga Rao result. The Shannon-Gallager-Berlekamp argument is similar to Blahut's in some ways, but it is less amenable to exact asymptotics. The Haroutunian argument is combinatorial and even farther removed from i.i.d. sums.

Blahut's argument proceeds as follows. Assume $R_\infty < R < C$ and let (f, φ) be an (N, R) code. Let $\{\mathcal{U}_m\}_{m \in \mathcal{M}}$ denote the decision regions of φ corresponding to each message $m \in \mathcal{M}$. Let $Q \in \mathcal{P}(\mathcal{Y})$ be an auxiliary output distribution. Let $W(\mathbf{y}^N | \mathbf{x}^N) := \prod_{n=1}^N W(y_n | x_n)$ and $Q(\mathbf{y}^N) := \prod_{n=1}^N Q(y_n)$. Since $\sum_{\mathbf{y}^N \in \mathcal{Y}^N} Q(\mathbf{y}^N) = 1$ and $|\mathcal{M}| \geq e^{NR}$, there must be a message $m \in \mathcal{M}$ such that $Q\{\mathcal{U}_m\} \leq e^{-NR}$. Let $\mathbf{x}^N := f(m)$ be the codeword for this message. It is clear that $e(f, \varphi) \geq e_m(f, \varphi) = W\{\mathcal{U}_m^c | \mathbf{x}^N\}$.

Now consider the hypothesis test over the set \mathcal{Y}^N in which $W(\cdot | \mathbf{x}^N)$ is the null hypothesis (H_0) and the i.i.d. output distribution Q is the alternate hypothesis (H_1). One feasible test is to accept H_0 on \mathcal{U}_m and H_1 on \mathcal{U}_m^c , resulting in type-I and type-II error probabilities of $W(\mathcal{U}_m^c | \mathbf{x}^N) = e_m(f, \varphi)$ and $Q\{\mathcal{U}_m\}$, respectively. Since $\alpha_{W(\cdot | \mathbf{x}^N), Q}^*(NR)$ denotes the minimum type-I error probability, optimized over all tests, subject to the constraint that the type-II error probability does not exceed e^{-NR} (cf. (2)), we evidently must have

$$e(f, \varphi) \geq \alpha_{W(\cdot | \mathbf{x}^N), Q}^*(NR). \quad (4)$$

The error exponent of this test can be expressed via the following definition. For any $V \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$, $P \in \mathcal{P}(\mathcal{X})$ and $Q \in \mathcal{P}(\mathcal{Y})$, define $D(V||Q|P) := \sum_{x \in \mathcal{X}} P(x) D(V(\cdot|x)||Q)$.

Definition 3.1: For any $r \in \mathbb{R}_+$, $P \in \mathcal{P}(\mathcal{X})$ and $Q \in \mathcal{P}(\mathcal{Y})$

$$e_{\text{SP}}(Q, P, r) := \inf_{V \in \mathcal{P}(\mathcal{Y}|\mathcal{X}) : D(V||Q|P) \leq r} D(V||W|P). \quad (5)$$

Then the optimal type-I error exponent can be shown to be (e.g. [28, Section V]) $e_{\text{SP}}(Q, P, R)$, where P is the empirical distribution of \mathbf{x}^N .

Note that this exponent depends on the output distribution Q , which is to be selected. This distribution can be chosen to depend on P , since it can depend on the code, although allowing such dependence necessitates a restriction to constant composition codes. In the original argument [28, Section V], this freedom is not used, and Q depends on R (and the channel) but not P . Pre-factors aside, it is not clear that this choice yields the standard sphere-packing exponent when (5) is maximized over P . This is asserted to be the case in [28, Theorem 19] and [10, Theorem 10.1.4], but each of these proofs has a nontrivial gap⁴. Moreover, a numerical study indicates that for the Z-channel and for this choice of Q , $E_{\text{SP}}(R) < \max_P e_{\text{SP}}(Q, P, R)$, for a broad range of rates. For symmetric channels, Q can indeed be chosen independently of P [21], and so the code need not be constant composition. But in the general case, it appears that some dependence is necessary if one hopes to obtain the sphere-packing exponent.

Our choice of Q will depend on P and give the sphere-packing exponent. Thus, one of the ancillary contributions of this paper is to give a complete proof that the hypothesis testing reduction described can be used to obtain the sphere-packing exponent. In fact, using the hypothesis testing reduction, we shall prove the stronger result that the exponent on the error probability of any constant-composition code with composition P is upper bounded by $E_{\text{SP}}(R, P)$; previously, the only proof of this fact used combinatorial techniques.

It is worth noting that the Shannon-Gallager-Berlekamp proof also involves the choice of an output distribution. Their choice of output distribution also depends on P , but it is defined differently from ours. Our choice yields the $E_{\text{SP}}(R, P)$ exponent, whereas Shannon-Gallager-Berlekamp only establish an exponent of $E_{\text{SP}}(R)$.

Before concluding this section, it is instructive to consider a binary symmetric channel (BSC) with crossover probability $p \in (0, 1/2)$ in order to see why the slope related term arises in Theorem 2.1. One can check that the output distribution mentioned in [21, Eq. 9] reduces to the uniform distribution and for these particular choices,

$$\alpha_{W(\cdot|\mathbf{x}^N), Q}^*(NR) \geq \sum_{n=n_R^*+1}^N \binom{N}{n} p^n (1-p)^{N-n} = \Pr \left\{ \frac{1}{N} \sum_{n=1}^N Z_n \geq \frac{n_R^*+1}{N} \right\}, \quad (6)$$

where $\{Z_n\}_{n=1}^N$ are i.i.d. Bernoulli random variables with parameter p and n_R^* is the largest $k \in \mathbb{Z}^+$ satisfying

$$e^{-NR} \geq \sum_{n=0}^k \binom{N}{n} 2^{-N} = \Pr \left\{ \frac{1}{N} \sum_{n=1}^N \tilde{Z}_n \leq \frac{k}{N} \right\}, \quad (7)$$

⁴Specifically, the argument for [28, Theorem 19] seems to proceed as if Lagrange multipliers of $\max_P E_{\text{SP}}(R, P)$ and $\max_P e_{\text{SP}}(Q, P, R)$ are the same, which is not evident. For [10, Theorem 10.1.4], only $e_{\text{SP}}(Q, P_R^*, R) = \max_P E_{\text{SP}}(R, P)$ is shown, where P_R^* attains $\max_P E_{\text{SP}}(R, P)$, which does not imply the claim.

where $\{\tilde{Z}_n\}_{n=1}^N$ are i.i.d. Bernoulli random variables with parameter $1/2$. Provided that $k/N < 1/2$, one can apply Theorem 1.1 to the right side of (7) to have

$$\Pr \left\{ \frac{1}{N} \sum_{n=1}^N \tilde{Z}_n \leq \frac{k}{N} \right\} \geq \frac{K_1}{\sqrt{N}} e^{-ND(k/n||1/2)}, \quad (8)$$

where $D(k/n||1/2) := k/n \log \frac{k/n}{1/2} + (1 - k/n) \log \frac{1-k/n}{1/2}$ and K_1 is a positive constant. Plugging (8) into (7) and recalling the definition of n_R^* , one can verify that

$$\frac{n_R^*}{N} \leq h^{-1} \left(\log 2 - R + \frac{\log \sqrt{N}}{N} - \frac{\log K_1}{N} \right) \quad (9)$$

By plugging (9) into (6), applying Theorem 1.1 on the right side of (6) and carrying out the algebra, one can verify that

$$\alpha_{W(\cdot|X^N), Q}^*(NR) \geq \frac{K_2}{\sqrt{N}} e^{-NE_{\text{SP}}\left(R - \frac{\log \sqrt{N}}{N}\right)} \geq \frac{K_3}{N^{0.5(1+|E'_{\text{SP}}(R)|)}} e^{-NE_{\text{SP}}(R)}, \quad (10)$$

where K_2, K_3 are positive constants and the last inequality follows by expanding $E_{\text{SP}}(\cdot)$ as a power series about R . Note that if $\frac{n_R^*}{N}$ were constant in N , then applying Theorem 1.1 to (6) would give a pre-factor with an order of $1/\sqrt{N}$. But Eq. (9) shows that $\frac{n_R^*}{N}$ increases with N at a rate of $(\log N)/N$. While this increase is too slow to affect the exponent, it does affect the order of the pre-factor.

Finally, note that the arguments leading to (10) are nothing but the “packing of Hamming spheres”. To be specific, one can check that (e.g. [2]) for this channel, the error probability of any (N, R) code is lower bounded by that of a hypothetical “sphere-packed code” with the same parameters. A sphere-packed code is a code such that the decoding region of each codeword is an Hamming sphere of a certain radius, say $\lceil N\delta(R) \rceil$ with $\delta(R) > 0$, possibly excluding some strings in the outermost layer and the union of these spheres equals $\{0, 1\}^N$. For the sphere-packed code, an error occurs when the noise pushes the received signal outside of the Hamming ball of radius n_R^* centered at the codeword, whose probability is precisely the right side of (6). By employing the upper bound given in (9), one can deduce (10).

By continuing this sphere-packing analogy, one can intuitively view the lower bound obtained via the hypothesis testing reduction as the error probability of a hypothetical sphere-packed (N, R) code on \mathcal{Y}^N with $\log \frac{Q(\cdot)}{W(\cdot|X^N)}$ used instead of Hamming distance. Note that the extra term in the pre-factor essentially stems from the approximation of the “maximal packing radius” of the spheres under this metric.

B. Selecting the output distribution

In order to describe our output distribution, we require the following technical results.

For any $Q \in \mathcal{P}(\mathcal{Y})$ and $\lambda \in [0, 1]$, define

$$\Lambda_{Q,P}(\lambda) := \begin{cases} E_P \left[\log E_{W(\cdot|X)} \left[\left(\frac{Q(Y)}{W(Y|X)} \right)^\lambda \right] \right], & \lambda \in (0, 1), \\ 0, & \lambda = 0. \end{cases}$$

For any $R \in \mathbb{R}^+$, define

$$\mathcal{P}_R(\mathcal{X}) := \{P \in \mathcal{P}(\mathcal{X}) : E_{\text{SP}}(R, P) > 0\}, \quad (11)$$

$$\mathcal{P}_{P,W}(\mathcal{Y}) := \{Q \in \mathcal{P}(\mathcal{Y}) : \forall x \in \mathcal{S}(P), \mathcal{S}(Q) \cap \mathcal{S}(W(\cdot|x)) \neq \emptyset\}, \quad (12)$$

$$\tilde{\mathcal{P}}_{P,W}(\mathcal{Y}) := \{Q \in \mathcal{P}(\mathcal{Y}) : \forall x \in \mathcal{S}(P), Q \gg W(\cdot|x)\}. \quad (13)$$

Further, given any $R > R_\infty$ and $P \in \mathcal{P}(\mathcal{X})$,

$$K_{R,P} : \mathbb{R}_+ \times \mathcal{P}_{P,W}(\mathcal{Y}) \rightarrow \mathbb{R}, \text{ s.t. } K_{R,P}(\rho, Q) = -\rho R - (1 + \rho)\Lambda_{Q,P}(\rho/(1 + \rho)), \quad (14)$$

for all $(\rho, Q) \in \mathbb{R}_+ \times \mathcal{P}_{P,W}(\mathcal{Y})$.

Proposition 3.1: (Saddle-point) Consider any $R_\infty < R < C$ and $P \in \mathcal{P}_R(\mathcal{X})$.

- (i) $K_{R,P}(\cdot, \cdot)$ has a saddle-point with the saddle-value $E_{\text{SP}}(R, P)$.
- (ii) Any saddle-point of $K_{R,P}(\cdot, \cdot)$, say (ρ^*, Q^*) , satisfies $(\rho^*, Q^*) \in \mathbb{R}^+ \times \tilde{\mathcal{P}}_{P,W}(\mathcal{Y})$.

Proof: The proof is provided in the Appendix A. ■

Let $S(R, P)$ denote the set of saddle-points of $K_{R,P}(\cdot, \cdot)$. Moreover,

$$S(R, P)|_{\mathbb{R}_+} := \{\rho \in \mathbb{R}_+ : \exists Q \in \mathcal{P}_{P,W}(\mathcal{Y}), \text{ s.t. } (\rho, Q) \in S(R, P)\}, \quad (15)$$

$$S(R, P)|_{\mathcal{P}_{P,W}(\mathcal{Y})} := \{Q \in \mathcal{P}_{P,W}(\mathcal{Y}) : \exists \rho \in \mathbb{R}_+, \text{ s.t. } (\rho, Q) \in S(R, P)\}, \quad (16)$$

Proposition 3.2: (Uniqueness of the saddle-point) For any $R_\infty < R < C$ and $P \in \mathcal{P}_R(\mathcal{X})$, $S(R, P)$ is a singleton.

Proof: The proof is given in the Appendix B. ■

Definition 3.2: Fix any $R_\infty < R < C$.

$$\rho_{R,\cdot}^* : \mathcal{P}_R(\mathcal{X}) \rightarrow \mathbb{R}_+, \text{ s.t. } \rho_{R,P}^* = S(R, P)|_{\mathbb{R}_+}, \quad (17)$$

$$Q_{R,\cdot}^* : \mathcal{P}_R(\mathcal{X}) \rightarrow \mathcal{P}_{P,W}(\mathcal{Y}), \text{ s.t. } Q_{R,P}^* = S(R, P)|_{\mathcal{P}_{P,W}(\mathcal{Y})}. \quad (18)$$

Observe that owing to Proposition 3.2, both (17) and (18) are well-defined. The distribution $Q_{R,\cdot}^*$ in (18) will be our output distribution.

Proposition 3.3: (Differentiability of $E_{\text{SP}}(\cdot, P)$) Consider any $R_\infty < R < C$ and $P \in \mathcal{P}_R(\mathcal{X})$. $E_{\text{SP}}(\cdot, P)$ is differentiable with $\rho_{R,P}^* = -\left.\frac{\partial E_{\text{SP}}(r, P)}{\partial r}\right|_{r=R}$.

Proof: The proof is given in the Appendix C. ■

Proposition 3.4: (Continuity of the saddle-point) Consider any $R_\infty < R < C$. Both $\rho_{R,\cdot}^*$ and $Q_{R,\cdot}^*$ are continuous on $\mathcal{P}_R(\mathcal{X})$.

Proof: The proof is provided in the Appendix D. ■

For any $R_\infty < R < C$ and $P \in \mathcal{P}_R(\mathcal{X})$, let $e_{\text{SP}}(R, P, r) := e_{\text{SP}}(Q_{P,R}^*, P, r)$ and $e_{\text{SP}}(R, P) := e_{\text{SP}}(R, P, R)$.

Theorem 3.1: (Equality of the exponents) For any $R_\infty < R < C$ and $P \in \mathcal{P}_R(\mathcal{X})$,

$$e_{\text{SP}}(R, P) = E_{\text{SP}}(R, P). \quad (19)$$

Proof: The proof is given in the Appendix E. ■

Remark 3.1: Recalling the discussion in the previous section, the equality of the exponents theorem, i.e. Theorem 3.1, ensures that the exponent of the lower bound on the error probability emerging as a result of binary hypothesis testing reduction in which $Q_{R,\cdot}^*$ is the alternate distribution matches the sphere-packing exponent. \diamond

C. Hypothesis testing reduction

For any $\nu, R \in \mathbb{R}^+$, define $\mathcal{P}_{R,\nu}(\mathcal{X}) := \{P \in \mathcal{P}(\mathcal{X}) : E_{\text{SP}}(R, P) \geq \nu\}$. Fix some $R \in (R_\infty, C)$ and some sufficiently small $\nu > 0$ that only depends on W and R . Application of the hypothesis testing reduction of Section III-A to an (N, R) constant composition code (f, φ) with common composition⁵ $P \in \mathcal{P}_{R,\nu}(\mathcal{X})$ by using $Q_{R,P}^*$ as the auxiliary output distribution yields (recall (4))

$$e(f, \varphi) \geq \alpha_N(R), \quad (20)$$

where $\alpha_N(R) := \alpha_{W(\cdot|\mathbf{x}^N), Q_{R,P}^*}(NR)$. On account of (20), in order to lower bound the maximal error probability of our code, it suffices to evaluate $\alpha_N(R)$.

However, since $Q_{R,P}^* \gg W(\cdot|\mathbf{x}^N)$ (cf. item (ii) of the saddle-point proposition, i.e. Proposition 3.1), but not necessarily⁶ $Q_{R,P}^* \equiv W(\cdot|\mathbf{x}^N)$, we need to do little more work. To this end, we define

$$\tilde{\mathcal{T}}(Q, \hat{Q}) := \left\{ T \in \mathcal{T}(Q, \tilde{Q}) : \mathcal{U}_T \cap [\mathcal{S}(\hat{Q}) \setminus \mathcal{S}(Q, \hat{Q})] = \emptyset \text{ and } \mathcal{U}_T^c \cap [\mathcal{S}(Q) \setminus \mathcal{S}(Q, \hat{Q})] = \emptyset \right\}, \quad (21)$$

where $\mathcal{S}(Q, \hat{Q}) := \mathcal{S}(Q) \cap \mathcal{S}(\hat{Q})$.

The proof of the following result is straightforward.

Lemma 3.1: For any $r \in \mathbb{R}^+$,

$$\alpha_{Q, \hat{Q}}^*(r) = \min_{T \in \tilde{\mathcal{T}}(Q, \hat{Q}) : \beta_T \leq e^{-r}} \alpha_T. \quad (22)$$

Lemma 3.2: For any $T \in \tilde{\mathcal{T}}(Q, \hat{Q})$, we have

$$\alpha_T = Q \left\{ \mathcal{S}(Q, \tilde{Q}) \right\} Q \left\{ \mathcal{U}_T^c \mid \mathcal{S}(Q, \hat{Q}) \right\}, \quad \beta_T = \hat{Q} \left\{ \mathcal{S}(Q, \hat{Q}) \right\} \hat{Q} \left\{ \mathcal{U}_T \mid \mathcal{S}(Q, \hat{Q}) \right\}, \quad (23)$$

where the conditional probabilities are induced by Q and \hat{Q} , respectively.

Proof: The result is obvious from the law of total probability and recalling (21). ■

Observe that owing to (22) we have⁷

$$\alpha_N(R) = \min_{T \in \tilde{\mathcal{T}}(W(\cdot|\mathbf{x}^N), Q_{R,P}^*) : \beta_T \leq e^{-NR}} \alpha_T. \quad (24)$$

In order to apply Lemmas 3.1 and 3.2 to our particular case, we need the following definition.

Definition 3.3: Given any $C > R > R_\infty$ and $P \in \mathcal{P}_R(\mathcal{X})$,

$$W_{R,P}^-(\cdot|x) := \begin{cases} \tilde{W}_{1^-, Q_{R,P}^*}(\cdot|x), & \text{if } x \in \mathcal{S}(P), \\ W(\cdot|x), & \text{else,} \end{cases} \quad (25)$$

⁵If $P \in \mathcal{P}_{R,\nu}(\mathcal{X})^c$, then it is possible to prove that (3) is true. See Lemma F.3 in the Appendix F.

⁶We have this equivalence if we consider a positive channel, for example.

⁷ $\tilde{\mathcal{T}}(W(\cdot|\mathbf{x}^N), Q_{R,P}^*)$ is defined as in (21).

where

$$\tilde{W}_{1-,Q_{R,P}^*}(\cdot|x) := \lim_{\lambda \uparrow 1} \tilde{W}_{\lambda,Q_{R,P}^*}(\cdot|x), \forall x \in \mathcal{S}(P). \quad (26)$$

and $\tilde{W}_{\lambda,Q_{R,P}^*}(\cdot|x)$ is the *tilted distribution* as defined in (117) in the Appendix A.

Remark 3.2: One can check that for any $x \in \mathcal{S}(P)$,

$$\tilde{W}_{1-,Q_{R,P}^*}(\cdot|x) = \begin{cases} \frac{Q_{R,P}^*(y)}{Q_{R,P}^*\{\mathcal{S}(W(\cdot|x))\}}, & \text{if } y \in \mathcal{S}(W(\cdot|x)), \\ 0, & \text{else.} \end{cases} \quad (27)$$

Equation (27) and the fact that $Q_{R,P}^* \gg W(\cdot|x)$, for all $x \in \mathcal{S}(P)$, ensure that (25) is a well-defined stochastic matrix from \mathcal{X} to \mathcal{Y} . Moreover, it is clear that $W_{R,P}^-(\cdot|x) \equiv W(\cdot|x)$, for all $x \in \mathcal{X}$. \diamond

Returning to our application, since $Q_{R,P}^* \gg W(\cdot|\mathbf{x}^N)$, (23) implies that for any $\tilde{T} \in \tilde{\mathcal{T}}(W(\cdot|\mathbf{x}^N), Q_{R,P}^*)$, we have

$$\alpha_{\tilde{T}} = W\{\mathcal{U}_{\tilde{T}}^c|\mathbf{x}^N\}, \quad \beta_{\tilde{T}} = Q_{R,P}^*\{\mathcal{S}(W(\cdot|\mathbf{x}^N))\} W_{R,P}^-\{\mathcal{U}_{\tilde{T}}|\mathbf{x}^N\}, \quad (28)$$

where $W_{R,P}^-(\mathbf{y}^N|\mathbf{x}^N) := \prod_{n=1}^N W_{R,P}^-(y_n|x_n)$ and $W_{R,P}^-$ is defined in (25).

Also,

$$\log Q_{R,P}^*\{\mathcal{S}(W^N(\cdot|\mathbf{x}^N))\} = \sum_{n=1}^N \log Q_{R,P}^*\{\mathcal{S}(W(\cdot|x_n))\} \quad (29)$$

$$\begin{aligned} &= N \sum_{x \in \mathcal{S}(P)} P(x) \log Q_{R,P}^*\{\mathcal{S}(W(\cdot|x))\} \\ &= -ND(W_{R,P}^-||Q_{R,P}^*|P), \end{aligned} \quad (30)$$

where (29) follows since $\mathcal{S}(W(\cdot|\mathbf{x}^N)) = \mathcal{S}(W(\cdot|x_1)) \times \dots \times \mathcal{S}(W(\cdot|x_N))$ and (30) follows by noting

$$\log Q_{R,P}^*\{\mathcal{S}(W(\cdot|x))\} = -D(W_{R,P}^-(\cdot|x)||Q_{R,P}^*),$$

which is a direct consequence of (25).

Combining (28) and (30), we conclude that for any $\tilde{T} \in \tilde{\mathcal{T}}(W(\cdot|\mathbf{x}^N), Q_{R,P}^*)$

$$[\beta_{\tilde{T}} \leq e^{-NR}] \iff [W_{R,P}^-\{\mathcal{U}_{\tilde{T}}|\mathbf{x}^N\} \leq e^{-Nr(R,P)}], \quad (31)$$

where

$$r(R,P) := R - D(W_{R,P}^-||Q_{R,P}^*|P). \quad (32)$$

Observe that the right side of (31) defines a non-trivial constraint only if $r(R,P) > 0$, which we establish next.

To this end, we first define the following set:

$$\tilde{\mathcal{P}}_{P,W}(\mathcal{Y}|\mathcal{X}) := \{V \in \mathcal{P}(\mathcal{Y}|\mathcal{X}) : \forall x \in \mathcal{S}(P), V(\cdot|x) \ll W(\cdot|x)\}. \quad (33)$$

Lemma 3.3: (Positivity of $r(R,P)$) Given any $R_\infty < R < C$ and $P \in \mathcal{P}_R(\mathcal{X})$,

- (i) $\forall V \in \tilde{\mathcal{P}}_{P,W}(\mathcal{Y}|\mathcal{X})$, $D(V||Q_{R,P}^*|P) = D(V||W_{R,P}^-|P) + D(W_{R,P}^-||Q_{R,P}^*|P)$.
- (ii) $r(R,P) > 0$.

Proof: The proof is given in the Appendix G. ■

Now, consider a binary hypothesis testing setup with the null hypothesis (resp. alternate hypothesis) $W(\cdot|\mathbf{x}^N)$ (resp. $W_{R,P}^-(\cdot|\mathbf{x}^N)$). Owing to (20), (24), (28) and (31), we deduce that

$$e(f, \varphi) \geq \tilde{\alpha}_N(r(R, P)) := \min_{T' \in \tilde{T}(W(\cdot|\mathbf{x}^N), W_{R,P}^-(\cdot|\mathbf{x}^N)) : \beta_{T'} \leq e^{-Nr(R, P)}} \alpha_{T'}. \quad (34)$$

On account of (34), in order to lower bound the maximal error probability of our constant composition code, it suffices to evaluate $\tilde{\alpha}_N(r(R, P))$. Instead of directly characterizing $\tilde{\alpha}_N(r(R, P))$, we give a lower bound on it by means of a test that is easier to analyze. In order to define this test, we need the following “shifted exponent”.

Definition 3.4: Given any $C > R > R_\infty$, $r \in \mathbb{R}_+$ and $P \in \mathcal{P}_R(\mathcal{X})$,

$$\tilde{e}_{\text{SP}}(R, P, r) := \inf_{V \in \mathcal{P}(\mathcal{Y}|\mathcal{X}) : D(V||W_{R,P}^-|P) \leq r} D(V||W|P). \quad (35)$$

Lemma 3.4: (Shifted exponent) For any $R > R_\infty$, $P \in \mathcal{P}_R(\mathcal{X})$ and $r > D(W_{R,P}^-||Q_{R,P}^*|P)$, we have

$$\tilde{e}_{\text{SP}}(R, P, r - D(W_{R,P}^-||Q_{R,P}^*|P)) = e_{\text{SP}}(R, P, r).$$

Proof: Fix an arbitrary $R > R_\infty$, $P \in \mathcal{P}_R(\mathcal{X})$ and $r > D(W_{R,P}^-||Q_{R,P}^*|P)$. Define $\tilde{r} := r - D(W_{R,P}^-||Q_{R,P}^*|P)$. Clearly, $\tilde{r} \in \mathbb{R}^+$. On account of the fact that $\tilde{e}_{\text{SP}}(R, P, \tilde{r}) \leq \tilde{e}_{\text{SP}}(R, P, 0) = D(W_{R,P}^-||W|P) < \infty$, it is easy to see that

$$\tilde{e}_{\text{SP}}(R, P, \tilde{r}) = \min_{V \in \tilde{\mathcal{P}}_{P,W}(\mathcal{Y}|\mathcal{X}) : D(V||W_{R,P}^-|P) \leq \tilde{r}} D(V||W|P). \quad (36)$$

Similarly,

$$e_{\text{SP}}(R, P, r) = \min_{V \in \tilde{\mathcal{P}}_{P,W}(\mathcal{Y}|\mathcal{X}) : D(V||Q_{R,P}^*|P) \leq r} D(V||W|P). \quad (37)$$

The item (i) of Lemma 3.3 ensures that the feasible regions of the right sides of (36) and (37) are the same. Since the cost functions of the two problems are the same, the lemma follows. \blacksquare

Fix an arbitrary $\zeta \in \mathbb{R}^+$ and let $\epsilon_N := (\frac{1}{2} + \zeta) \frac{\log N}{N}$ (resp. $\tilde{\epsilon}_N := \epsilon_N - \frac{1}{N}$) and define $R_N := R - \epsilon_N$ (resp. $\tilde{R}_N := R - \tilde{\epsilon}_N$). Note that for all sufficiently large $N \in \mathbb{Z}^+$, $C > \tilde{R}_N > R_N > R_\infty$. Throughout, we consider such an $N \in \mathbb{Z}^+$. Further, similar to (32), define $r_N(P, R) := R_N - D(W_{R,P}^-||Q_{R,P}^*|P)$ (resp. $\tilde{r}_N(P, R) := \tilde{R}_N - D(W_{R,P}^-||Q_{R,P}^*|P)$). Also,

$$A_N := \left\{ \mathbf{y}^N : \frac{1}{N} \sum_{n=1}^N \log \frac{W(y_n|x_n)}{W_{R,P}^-(y_n|x_n)} > r_N(R, P) - \tilde{e}_{\text{SP}}(R, P, r_N(R, P)) \right\}, \quad (38)$$

$$A_N^c = \left\{ \mathbf{y}^N : \frac{1}{N} \sum_{n=1}^N \log \frac{W_{R,P}^-(y_n|x_n)}{W(y_n|x_n)} \geq \tilde{e}_{\text{SP}}(R, P, r_N(R, P)) - r_N(R, P) \right\}. \quad (39)$$

Equations (38) and (39) are the decision regions of the test, i.e. the test decides $W(\cdot|\mathbf{x}^N)$ if $\mathbf{y}^N \in A_N$ and $W_{R,P}^-(\cdot|\mathbf{x}^N)$ if $\mathbf{y}^N \in A_N^c$. Let

$$\alpha_N := W\{A_N^c|\mathbf{x}^N\}, \quad \beta_N := W_{R,P}^-\{A_N|\mathbf{x}^N\}, \quad (40)$$

denote the error probabilities of the aforementioned test.

Remark 3.3: The analysis of the events A_N and A_N^c would be direct applications of Bahadur-Ranga Rao but for two complications: First, the random variables in the sum are not i.i.d., only independent. This does not present a major difficulty, as one can prove a version of the Bahadur-Ranga Rao result for independent random variables that is weaker but sufficient for our purposes, which is given in the next section. The second complication is that the threshold in both events depends on N . One could define constant-threshold versions of these events by replacing $r_N(R, P)$ with $r(R, P)$. Applying exact asymptotics to the resulting events would yield a lower bound on α_N of the order $\frac{1}{\sqrt{N}} \exp(-N E_{\text{SP}}(R, P))$ and show that β_N is of the order $\frac{1}{\sqrt{N}} \exp(-N r(R, P))$. The problem with this approach is that $e(f, \varphi)$ is lower bounded by the type-I error probability of the optimal test whose type-II probability does not exceed $e^{-N r(R, P)}$. From the above expression of β_N , we see that the aforementioned test is not optimal because, although it is a likelihood ratio test, it is “undershooting” the type-II constraint due to the $1/\sqrt{N}$ pre-factor. By replacing $r(R, P)$ with $r_N(R, P)$, we ensure that β_N does not undershoot the constraint (in fact, it will violate it by a small amount). The $r_N(R, P)$ fluctuations will give rise to the slope term in the pre-factor of the probability of A_N . \diamond

D. Analysis of the hypothesis test

In this section, we analyze the hypothesis test stated in the previous section. In order to accomplish this, we begin with the following generalization of the Bahadur-Ranga Rao theorem.

1) *Sharp Lower Bound:* The content of this section resembles Dembo-Zeitouni’s proof of Theorem 1.1 (cf. [26, Theorem 3.7.4]). Here, we essentially use the same ideas but generalize them to cover non-identical case.

Let $\{Z_i\}_{i=1}^n$ be a sequence of independent, real-valued random variables and λ_i be the law of Z_i . Assume $\sum_{i=1}^n \text{Var}[Z_i] \in \mathbb{R}^+$. Define $\Lambda_i(\delta) := \log \mathbb{E}[e^{\delta Z_i}]$, $M_i(\delta) := e^{\Lambda_i(\delta)}$ and the Fenchel-Legendre transform of $\frac{1}{n} \sum_{i=1}^n \Lambda_i(\cdot)$ as:

$$\forall x \in \mathbb{R}, \Lambda_n^*(x) := \sup_{\delta \in \mathbb{R}} \left\{ x\delta - \frac{1}{n} \sum_{i=1}^n \Lambda_i(\delta) \right\}. \quad (41)$$

Let $q \in \mathbb{R}$ be such that $\exists \eta \in (0, 1]$ with the following properties:

- (i) There exists a neighborhood of η such that $\frac{1}{n} \sum_{i=1}^n \Lambda_i(\delta) < \infty$, for all δ in this neighborhood.
- (ii) $\frac{1}{n} \sum_{i=1}^n \Lambda_i'(\eta) = q$.

Remark 3.4: The reason to choose 1 as an upper bound on η above is just for our application of the main result of this section in the sequel. One may use an arbitrary constant and modify the result accordingly. \diamond

Define $\hat{S}_n := \frac{1}{n} \sum_{i=1}^n Z_i$ and let μ_n denote the law of \hat{S}_n . Also, define the probability measure $\tilde{\lambda}_i$ via

$$\frac{d\tilde{\lambda}_i}{d\lambda_i}(z_i) := e^{\eta z_i - \Lambda_i(\eta)}. \quad (42)$$

Let $\tilde{\mu}_n$ denote the law of \hat{S}_n when Z_i are independent with the marginal law $\tilde{\lambda}_i$. Further, define⁸ $T_i := Z_i - \mathbb{E}_{\tilde{\lambda}_i}[Z_i]$, $m_{2,n} := \sum_{i=1}^n \text{Var}_{\tilde{\lambda}_i}[T_i]$, $m_{3,n} := \sum_{i=1}^n \mathbb{E}_{\tilde{\lambda}_i}[|T_i|^3]$ and $W_n := \frac{1}{\sqrt{m_{2,n}}} \sum_{i=1}^n T_i$. Also, $K_n(\eta) := \frac{15\sqrt{2\pi}m_{3,n}}{m_{2,n}}$.

⁸We shall show that all of the following quantities are well-defined in the proof of the following proposition, given in the Appendix H.

Proposition 3.5: (Sharp lower bound) Provided that

$$\sqrt{m_{2,n}} \geq 1 + (1 + K_n(\eta))^2 \quad (43)$$

holds,

$$\mu_n([q, \infty)) \geq e^{-n\Lambda_n^*(q)} \frac{e^{-K_n(\eta)}}{2\sqrt{2\pi m_{2,n}}}. \quad (44)$$

Proof: The proof is given in the Appendix H. ■

2) *Analysis of α_N and β_N :* In this section, we apply Proposition 3.5 to lower bound α_N and β_N given in (40).

To this end, we begin with the following technical results.

Definition 3.5: Let $C > R > R_\infty$ and $P \in \mathcal{P}_R(\mathcal{X})$ be arbitrary but fixed. Let $\lambda \in \mathbb{R}$ be arbitrary.

$$\Lambda_{0,P,x}(\lambda) := \log \mathbb{E}_{W(\cdot|x)} \left[e^{\lambda \log \frac{W_{R,P}^-(Y|x)}{W(Y|x)}} \right], \quad (45)$$

$$\Lambda_{0,P}(\lambda) := \sum_{x \in \mathcal{X}} P(x) \Lambda_{0,P,x}(\lambda), \quad (46)$$

$$\Lambda_{1,P,x}(\lambda) := \log \mathbb{E}_{W_{R,P}^-(\cdot|x)} \left[e^{\lambda \log \frac{W(Y|x)}{W_{R,P}^-(Y|x)}} \right], \quad (47)$$

$$\Lambda_{1,P}(\lambda) := \sum_{x \in \mathcal{X}} P(x) \Lambda_{1,P,x}(\lambda), \quad (48)$$

Remark 3.5:

- (i) Since $W_{R,P}^-(\cdot|x) \equiv W(\cdot|x)$ for all $x \in \mathcal{X}$, each quantity given in Definition 3.5 is well-defined. Also, one can check that $\Lambda_{1,P,x}(\lambda) = \Lambda_{0,P,x}(1 - \lambda)$, which, in turn, implies that $\Lambda_{1,P}(\lambda) = \Lambda_{0,P}(1 - \lambda)$.
- (ii) The fact that $W_{R,P}^-(\cdot|x) \equiv W(\cdot|x)$ for all $x \in \mathcal{X}$ also ensures that $\Lambda_{0,P}(\lambda), \Lambda_{1,P}(\lambda) \in \mathbb{R}$ and hence $\Lambda_{0,P}(\cdot), \Lambda_{1,P}(\cdot) \in \mathcal{C}^\infty(\mathbb{R})$.
- (iii) Consider any $\lambda \in \mathbb{R}$. It is easy to verify the following (for the sake of notational convenience, we denote partial derivatives with respect to λ as the ordinary ones):

$$\Lambda'_{0,P,x}(\lambda) = \mathbb{E}_{\tilde{W}_{\lambda,P}(\cdot|x)} \left[\log \frac{W_{R,P}^-(Y|x)}{W(Y|x)} \right], \quad (49)$$

$$\Lambda'_{0,P}(\lambda) = \sum_{x \in \mathcal{X}} P(x) \Lambda'_{0,P,x}(\lambda), \quad (50)$$

$$\Lambda''_{0,P,x}(\lambda) = \text{Var}_{\tilde{W}_{\lambda,P}(\cdot|x)} \left[\log \frac{W_{R,P}^-(Y|x)}{W(Y|x)} \right], \quad (51)$$

$$\Lambda''_{0,P}(\lambda) = \sum_{x \in \mathcal{X}} P(x) \Lambda''_{0,P,x}(\lambda), \quad (52)$$

where $\tilde{W}_{\lambda,P}(\cdot|x) := \tilde{W}_{\lambda, W_{R,P}^-(\cdot|x)}$ (cf. (117)) for the sake of notational convenience.

Further, item (ii) above ensures that

$$\Lambda'_{1,P,x}(\lambda) = -\Lambda'_{0,P,x}(1 - \lambda), \quad \Lambda'_{1,P}(\lambda) = -\Lambda'_{0,P}(1 - \lambda), \quad (53)$$

$$\Lambda''_{1,P,x}(\lambda) = \Lambda''_{0,P,x}(1 - \lambda), \quad \Lambda''_{1,P}(\lambda) = \Lambda''_{0,P}(1 - \lambda), \quad (54)$$

for any $\lambda \in \mathbb{R}$.

(iv) We have

$$\Lambda'_{0,P}(0) = -\Lambda'_{1,P}(1) = -D(W||W_{R,P}^-|P), \quad (55)$$

$$\Lambda'_{0,P}(1) = -\Lambda'_{1,P}(0) = D(W_{R,P}^-||W|P), \quad (56)$$

as a direct consequence of (50) and (53). \diamond

Lemma 3.5: (Positive variance) Let $C > R > R_\infty$ and $P \in \mathcal{P}_R(\mathcal{X})$ be arbitrary. For all $\lambda \in [0, 1]$, $\Lambda''_{0,P}(\lambda) > 0$.

Proof: Consider any $C > R > R_\infty$, $P \in \mathcal{P}_R(\mathcal{X})$ and recall that $r(R, P) = R - D(W_{R,P}^-||Q_{R,P}^*|P)$ (cf. (32)).

For contradiction, suppose there exists $\lambda \in [0, 1]$ such that $\Lambda''_{0,P}(\lambda) = 0$. We have

$$[\Lambda''_{0,P}(\lambda) = 0] \iff \left[\forall x \in \mathcal{S}(P), \log \frac{W_{R,P}^-(Y|x)}{W(Y|x)} = \Lambda'_{0,P,x}(\lambda), W(\cdot|x) - (\text{a.s.}) \right] \quad (57)$$

$$\iff \left[\forall x \in \mathcal{S}(P), W(Y|x) = W_{R,P}^-(Y|x)e^{-\Lambda'_{0,P,x}(\lambda)}, W(\cdot|x) - (\text{a.s.}) \right], \quad (58)$$

where (57) follows from (49), (51) and (52). Summing the right side of (58) over $y \in \mathcal{S}(W(\cdot|x))$ yields

$$\forall x \in \mathcal{S}(P), \Lambda'_{0,P,x}(\lambda) = 0. \quad (59)$$

Combining (58) and (59) and recalling the definition of $W_{R,P}^-$ (cf. (25)), we deduce that

$$[\Lambda''_{0,P}(\lambda) = 0] \iff \left[\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, W(y|x) = W_{R,P}^-(y|x) \right]. \quad (60)$$

The right side of (60) implies that $\tilde{e}_{\text{SP}}(R, P, r) = 0$ for all $r \in \mathbb{R}_+$ and in particular $\tilde{e}_{\text{SP}}(R, P, r(R, P)) = 0$.

This observation, coupled with the equality of the exponents theorem, i.e. Theorem 3.1, and the shifted exponent exponent lemma, i.e. Lemma 3.4, implies that $E_{\text{SP}}(R, P) = 0$ that contradicts the fact that $P \in \mathcal{P}_R(\mathcal{X})$. \blacksquare

Definition 3.6: Let $C > R > R_\infty$ be arbitrary. For any $(\lambda, P) \in [0, 1] \times \mathcal{P}_R(\mathcal{X})$,

$$m_{0,3}(\lambda, P) := \sum_{x \in \mathcal{S}(P)} P(x) E_{\tilde{W}_{\lambda,P}(\cdot|x)} \left[\left| \log \frac{W_{R,P}^-(Y|x)}{W(Y|x)} - \Lambda'_{0,P,x}(\lambda) \right|^3 \right], \quad (61)$$

$$m_{1,3}(\lambda, P) := \sum_{x \in \mathcal{S}(P)} P(x) E_{\tilde{W}_{1-\lambda,P}(\cdot|x)} \left[\left| \log \frac{W(Y|x)}{W_{R,P}^-(Y|x)} - \Lambda'_{1,P,x}(\lambda) \right|^3 \right]. \quad (62)$$

Note that owing to (53), (61) and (62), one can verify that

$$\forall (\lambda, P) \in [0, 1] \times \mathcal{P}_R(\mathcal{X}), m_{0,3}(\lambda, P) = m_{1,3}(1 - \lambda, P). \quad (63)$$

Lemma 3.6: (Continuity)

- (i) $\Lambda'_{0,\cdot}(\cdot)$ is continuous on $(0, 1] \times \mathcal{P}_R(\mathcal{X})$.
- (ii) $\Lambda''_{0,\cdot}(\cdot)$ is continuous on $(0, 1] \times \mathcal{P}_R(\mathcal{X})$.
- (iii) $m_{0,3}(\cdot, \cdot)$ is continuous on $(0, 1] \times \mathcal{P}_R(\mathcal{X})$.

(iv) $D(W_{R,\cdot}^- || Q_{R,\cdot}^* | \cdot)$ is continuous on $\mathcal{P}_R(\mathcal{X})$.

Proof: The proof is given in the Appendix I. ■

Lemma 3.7: Fix arbitrary $C > R > R_\infty$ and $P \in \mathcal{P}_R(\mathcal{X})$. For any $r \in \mathbb{R}^+$, we have

$$\tilde{e}_{\text{SP}}(R, P, r) = \max_{s \in \mathbb{R}_+} \{-sr + e_0(s, P)\}, \quad (64)$$

where

$$e_0(s, P) := -(1+s) \sum_{x \in \mathcal{S}(P)} P(x) \log \sum_{y \in \mathcal{S}(W(\cdot|x))} W(y|x)^{1/(1+s)} W_{R,P}^-(y|x)^{s/(1+s)} \quad (65)$$

for any $s \in \mathbb{R}_+$.

Proof: We have,

$$\begin{aligned} \tilde{e}_{\text{SP}}(R, P, r) &= \inf_{V \in \mathcal{P}(\mathcal{Y}|\mathcal{X}) : D(V || W_{R,P}^- | P) \leq r} D(V || W | P) \\ &= \max_{s \in \mathbb{R}_+} \min_{V \in \mathcal{P}(\mathcal{Y}|\mathcal{X})} \left\{ D(V || W | P) + s(D(V || W_{R,P}^- | P) - r) \right\} \end{aligned} \quad (66)$$

$$= \max_{s \in \mathbb{R}_+} \left\{ -sr + \sum_{x \in \mathcal{S}(P)} P(x) \min_{V(\cdot|x) \in \mathcal{P}(\mathcal{Y})} \left[D(V(\cdot|x) || W(\cdot|x)) + sD(V(\cdot|x) || W_{R,P}^-(\cdot|x)) \right] \right\} \quad (67)$$

$$= \max_{s \in \mathbb{R}_+} \{-sr + e_0(s, P)\}, \quad (68)$$

where (66) follows since Slater's condition holds (cf. [30, Corollary 28.2.1]), (68) follows by noting that

$$V_{P,s}^*(y|x) := \frac{W(y|x)^{1/(1+s)} W_{R,P}^-(y|x)^{s/(1+s)}}{\sum_{\tilde{y} \in \mathcal{Y}} W(\tilde{y}|x)^{1/(1+s)} W_{R,P}^-(\tilde{y}|x)^{s/(1+s)}},$$

attains the minimum in (67) for any $x \in \mathcal{S}(P)$ and recalling (65). ■

Corollary 3.1: Consider any $C > R > R_\infty$, $P \in \mathcal{P}_R(\mathcal{X})$. For all $r \in \mathbb{R}^+$, the set of maximizers of (64) is exactly $\partial \tilde{e}_{\text{SP}}(R, P, \cdot)(r)$.

Proof: Proof follows exactly the same lines as that of Lemma C.2. ■

Lemma 3.8: (Differentiability of the shifted exponent) Let $C > R > R_\infty$ and $r \in \mathbb{R}^+$ be given.

$$s^*(R, \cdot, r) : \mathcal{P}_R(\mathcal{X}) \rightarrow \mathbb{R}_+, \text{ s.t. } s^*(R, P, r) := -\frac{\partial \tilde{e}_{\text{SP}}(R, P, r)}{\partial r}, \forall P \in \mathcal{P}_R(\mathcal{X}), \quad (69)$$

is a well-defined function.

Proof: Consider any $P \in \mathcal{P}_R(\mathcal{X})$. For any $s \in \mathbb{R}_+$, (65), (45), (46), (50) and (52) imply that

$$\frac{\partial^2 e_0(s, P)}{\partial s^2} = -\frac{1}{(1+s)^3} \Lambda_{0,P}'' \left(\frac{s}{1+s} \right) < 0. \quad (70)$$

where the inequality follows from the positive variance lemma, i.e. Lemma 3.5. Equation (70) ensures the strict concavity of the cost function of (64) and hence the uniqueness of the maximizer. Recalling Corollary 3.1, this implies that (69) is well-defined. ■

The shifted exponent lemma, i.e. Lemma 3.4, and the differentiability of the shifted exponent, i.e. Lemma 3.8, immediately implies the following result.

Corollary 3.2: Given any $C > R > R_\infty$ and $P \in \mathcal{P}_R(\mathcal{X})$ and ,

$$\left. \frac{\partial \text{esp}(R, P, \tilde{r})}{\partial \tilde{r}} \right|_{\tilde{r}=r} = -s^*(R, P, r - \text{D}(W_{R,P}^- || Q_{R,P}^* | P)),$$

for any $r > \text{D}(W_{R,P}^- || Q_{R,P}^* | P)$.

Throughout this section, unless stated otherwise, suppose $C(W) > R > R_\infty$ and $P \in \mathcal{P}_R(\mathcal{X})$ be arbitrary and fixed.

Definition 3.7: Consider any $C > R > R_\infty$ and $P \in \mathcal{P}_R(\mathcal{X})$. Given any $z \in \mathbb{R}$,

$$\Lambda_{0,P}^*(z) := \sup_{\lambda \in \mathbb{R}} \{\lambda z - \Lambda_{0,P}(\lambda)\}, \quad (71)$$

$$\Lambda_{1,P}^*(z) := \sup_{\lambda \in \mathbb{R}} \{\lambda z - \Lambda_{1,P}(\lambda)\}. \quad (72)$$

Lemma 3.9: (Regularity) Fix any $C > R > R_\infty$ and $P \in \mathcal{P}_R(\mathcal{X})$. For any $0 < r < \text{D}(W || W_{R,P}^- | P)$,

- (i) $\Lambda_{0,P}^*(\tilde{\text{esp}}(R, P, r) - r) = \tilde{\text{esp}}(R, P, r)$.
- (ii) $\Lambda_{1,P}^*(r - \tilde{\text{esp}}(R, P, r)) = r$.
- (iii) There exists a unique $\eta(R, P, r) \in (0, 1)$, such that $\Lambda_{0,P}'(\eta(R, P, r)) = \tilde{\text{esp}}(R, P, r) - r$. In particular, $\eta(R, P, r) = \frac{s^*(R, P, r)}{1 + s^*(R, P, r)}$.

Proof: The proof is given in the Appendix J. ■

Next, we claim that

$$0 < r(R, P) < \text{I}(P; W) - \text{D}(W_{R,P}^- || Q_{R,P}^* | P) \leq \text{D}(W || W_{R,P}^- | P). \quad (73)$$

The first inequality follows from the positivity of $r(R, P)$ lemma, i.e. Lemma 3.3. The second inequality is clear from the definition of $r(R, P)$ and the fact that $P \in \mathcal{P}_R(\mathcal{X})$. The last inequality follows by noting

$$\text{D}(W_{R,P}^- || Q_{R,P}^* | P) + \text{D}(W || W_{R,P}^- | P) = \text{D}(W || Q_{R,P}^* | P) \geq \min_{Q \in \mathcal{P}(\mathcal{Y})} \text{D}(W || Q | P) = \text{I}(P; W),$$

where the first equality follows from the item (i) of Lemma 3.3 and the last one follows from (143). Hence, (73) follows.

Further, define

$$\Upsilon(W, R, \nu) := \max_{P \in \mathcal{P}_{R,\nu}(\mathcal{X})} \text{D}(W || Q_{R,P}^* | P), \quad H := \left[\frac{\frac{\nu}{2\Upsilon(W, R, \nu)}}{1 + \frac{\nu}{2\Upsilon(W, R, \nu)}}, 1 \right].$$

Since $\text{E}_{\text{Sp}}(\cdot, \cdot)$ is continuous (cf. Lemma F.2), $\mathcal{P}_{R,\nu}$ is closed and therefore, by noting the boundedness of $\mathcal{P}(\mathcal{X})$, is compact. Further, owing to the continuity of $\text{D}(W || Q_{R,\cdot}^* | \cdot)$ (cf. the item (iv) of the continuity lemma, i.e. Lemma 3.6) and the compactness of $\mathcal{P}_{R,\nu}(\mathcal{X})$, $\Upsilon(W, R, \nu)$ is well-defined and finite.

Lemma 3.10: For any $P \in \mathcal{P}_{R,\nu}(\mathcal{X})$

$$\eta(R, P, r) \in H, \quad \forall r \in (0, r(P, R)]. \quad (74)$$

Proof: Let $P \in \mathcal{P}_{R,\nu}(\mathcal{X})$ be arbitrary. Owing to the item (iii) of the regularity lemma, i.e. Lemma 3.9, it suffices to prove that for all $r \in (0, r(P, R)]$

$$\eta(R, P, r) \geq \frac{\frac{\nu}{2\Upsilon(W, R, \nu)}}{1 + \frac{\nu}{2\Upsilon(W, R, \nu)}}. \quad (75)$$

Moreover, the fact that $\eta(R, P, r) = s^*(R, P, r)/(1 + s^*(R, P, r))$ (cf. item (iii) of Lemma 3.9), (73), the convexity and the non-increasing property of $\tilde{e}_{\text{SP}}(R, P, \cdot)$, it suffices to show (75) for $r = r(R, P)$. The differentiability of the shifted exponent lemma, i.e. Lemma 3.8, and Corollary 3.2 imply that

$$s^*(R, P, r(R, P)) = - \left. \frac{\partial \tilde{e}_{\text{SP}}(R, P, r)}{\partial r} \right|_{r=r(R, P)} = - \left. \frac{\partial e_{\text{SP}}(R, P, r)}{\partial r} \right|_{r=R}. \quad (76)$$

Moreover, using the convexity and the non-increasing property of $e_{\text{SP}}(R, P, \cdot)$, one can see that

$$- \left. \frac{\partial e_{\text{SP}}(R, P, r)}{\partial r} \right|_{r=R} \geq \frac{\nu}{2} \frac{1}{(e_{\text{SP}}^{-1}(R, P, \cdot)(\nu/2) - R)} \geq \frac{\nu}{2\Upsilon(W, R, \nu)}, \quad (77)$$

where the last inequality follows by noting that $e_{\text{SP}}(R, P, r) = 0$ for all $r \geq D(W || Q_{R,P}^* | P)$. By combining (76) and (77), we deduce that

$$s^*(R, P, r(R, P)) \geq \frac{\nu}{2\Upsilon(W, R, \nu)}. \quad (78)$$

Since $\eta(R, P, r) = s^*(R, P, r)/(1 + s^*(R, P, r))$, (78) implies (74). ■

Finally, we define the following:

$$\overline{M}(\nu, W, R) := \max_{(\lambda, P) \in H \times \mathcal{P}_{R,\nu}} \frac{m_{0,3}(\lambda, P)}{\Lambda_{0,P}''(\lambda)}, \quad (79)$$

$$\overline{V}(\nu, W, R) := \max_{(\lambda, P) \in H \times \mathcal{P}_{R,\nu}} \Lambda_{0,P}''(\lambda), \quad (80)$$

$$\underline{V}(\nu, W, R) := \min_{(\lambda, P) \in H \times \mathcal{P}_{R,\nu}} \Lambda_{0,P}''(\lambda), \quad (81)$$

where H is as defined prior to Lemma 3.10. Recalling the compactness of H and $\mathcal{P}_{R,\nu}(\mathcal{X})$, the positive variance lemma, i.e. Lemma 3.5 and the continuity lemma, i.e. Lemma 3.6 ensures that (79), (80) and (81) are well-defined, positive and finite.

Define $K_{\max} := 2\sqrt{2\pi c} \overline{M}(\nu, W, R)$ with $c = 30/4$. Note that $K_{\max} \in \mathbb{R}^+$. Also, let $N \in \mathbb{Z}^+$ be sufficiently large, such that

$$\sqrt{N} \geq \frac{1 + (1 + K_{\max})^2}{\sqrt{\underline{V}(\nu, W, R)}}, \quad (82)$$

and consider such an N from now on.

Next, we apply the sharp lower bound proposition, i.e. Proposition 3.5, to α_N to deduce a lower bound. Observe that (51), (52) and the positive variance proposition, i.e. Proposition 3.5 and the item (iii) of the regularity lemma, i.e. Lemma 3.9 ensures the fulfillment of the assumptions under which Proposition 3.5 is stated. Moreover, (82) guarantees that (43) holds, and hence we can apply Proposition 3.5 to $W \{A_N^c | \mathbf{x}^N\}$ (cf. (39) and (40)) to deduce

$$\alpha_N \geq \frac{K}{\sqrt{N}} \exp\{-N \Lambda_{0,P}^*(\tilde{e}_{\text{SP}}(R, P, r_N(R, P)) - r_N(R, P))\}, \quad (83)$$

where we define

$$K := \frac{e^{-K_{\max}}}{2\sqrt{2\pi\bar{V}(\nu, W, R)}}. \quad (84)$$

Note that K only depends on W , R and ν .

Further, recalling the definition of β_N (cf. (38) and (40)) one can check that

$$\beta_N \geq W_{R,P}^- \left\{ \frac{1}{N} \sum_{n=1}^N \log \frac{W(Y_n|x_n)}{W_{R,P}^-(Y_n|x_n)} \geq \tilde{r}_N(R, P) - \tilde{e}_{\text{SP}}(R, P, \tilde{r}_N(R, P)) \mid \mathbf{x}^N \right\}. \quad (85)$$

Next, we apply the sharp lower bound proposition, i.e. Proposition 3.5, to the right side of (85) by noting the fact that the explanations provided prior to (83) are still valid (recall (53) and (54)) and infer the following

$$\beta_N \geq \frac{K}{\sqrt{N}} e^{-N\Lambda_{1,P}^*(\tilde{r}_N(R, P) - \tilde{e}_{\text{SP}}(R, P, \tilde{r}_N(R, P)))} = \frac{K}{\sqrt{N}} e^{-N\tilde{r}_N(R, P)} = \frac{KN^\zeta}{e} e^{-Nr(R, P)}, \quad (86)$$

where the first equality follows from the item (ii) of the regularity lemma, i.e. Lemma 3.9.

If we let $N \in \mathbb{Z}^+$ to be sufficiently large, so that

$$\frac{KN^\zeta}{e} > 1,$$

then (86) implies that $\beta_N > e^{-Nr(P, R)}$. Since our test is a likelihood ratio test, by violating the constraint we can only improve the optimal error performance, and hence (cf. (83))

$$\tilde{\alpha}_N(r(P, R)) \geq \alpha_N \geq \frac{K}{\sqrt{N}} e^{-N\Lambda_{0,P}^*(\tilde{e}_{\text{SP}}(R, P, r_N(R, P)) - r_N(R, P))},$$

which, in turn, implies that (cf. (34))

$$e(f, \varphi) \geq \frac{K}{\sqrt{N}} e^{-N\Lambda_{0,P}^*(\tilde{e}_{\text{SP}}(R, P, r_N(R, P)) - r_N(R, P))}. \quad (87)$$

E. Approximation of the exponent

In this final section, we approximate the exponent in (87) to conclude the proof.

To begin with, we note that (e.g. [26, Exercise 2.2.24]) $\Lambda_{0,P}^*(\cdot) \in \mathcal{C}^\infty(-D(W||W_{R,P}^-|P), D(W_{R,P}^-||W|P))$. Moreover, with the aid of the inverse function theorem and the item (iii) of the regularity lemma, i.e. Lemma 3.9, one can check that for any $r \in (0, D(W||W_{R,P}^-|P))$,

$$\Lambda_{0,P}^{*'}(\tilde{e}_{\text{SP}}(R, P, r) - r) = \eta(R, P, r), \quad \Lambda_{0,P}^{*''}(\tilde{e}_{\text{SP}}(R, P, r) - r) = \frac{1}{\Lambda_{0,P}''(\eta(R, P, r))}. \quad (88)$$

Define⁹

$$\delta(R, \nu, W) := R - \max_{P \in \mathcal{P}_{R,\nu}(\mathcal{X})} D(W_{R,P}^-||Q_{R,P}^*|P).$$

Observe that owing to Lemma 3.3, $\delta(R, \nu, W) > 0$. Hence, one can choose $N \in \mathbb{Z}^+$ to be sufficiently large, such that $\epsilon_N \leq \delta(R, \nu, W)/2$. Consider such an N from now on.

⁹Owing to the item (iv) of the continuity lemma, i.e. Lemma 3.6, and the compactness of $\mathcal{P}_{R,\nu}(\mathcal{X})$, the maximum is well-defined.

Using Taylor's theorem, for some $\bar{x} \in (\tilde{e}_{\text{SP}}(R, P, r_N(R, P)) - r_N(R, P), \tilde{e}_{\text{SP}}(R, P, r(R, P)) - r(R, P))$, we get

$$\begin{aligned}
\Lambda_{0,P}^*(\tilde{e}_{\text{SP}}(R, P, r_N(R, P)) - r_N(R, P)) &= \Lambda_{0,P}^*(\tilde{e}_{\text{SP}}(R, P, r(R, P)) - r(R, P)) + \{(\tilde{e}_{\text{SP}}(R, P, r_N(R, P)) - r_N(R, P)) \\
&\quad - (\tilde{e}_{\text{SP}}(R, P, r(R, P)) - r(R, P))\} \Lambda_{0,P}^{*'}(\tilde{e}_{\text{SP}}(R, P, r(R, P)) - r(R, P)) \\
&\quad + \frac{\Lambda_{0,P}^{*''}(\bar{x}) \{(\tilde{e}_{\text{SP}}(R, P, r_N(R, P)) - r_N(R, P)) - (\tilde{e}_{\text{SP}}(R, P, r(R, P)) - r(R, P))\}^2}{2} \\
&= \Lambda_{0,P}^*(\tilde{e}_{\text{SP}}(R, P, r(R, P)) - r(R, P)) + \epsilon_N \Lambda_{0,P}^{*'}(\tilde{e}_{\text{SP}}(R, P, r(R, P)) - r(R, P)) \\
&\quad + [\tilde{e}_{\text{SP}}(R, P, r_N(R, P)) - \tilde{e}_{\text{SP}}(R, P, r(R, P))] \Lambda_{0,P}^{*'}(\tilde{e}_{\text{SP}}(R, P, r(R, P)) - r(R, P)) \\
&\quad + \frac{\Lambda_{0,P}^{*''}(\bar{x}) \{(\tilde{e}_{\text{SP}}(R, P, r_N(R, P)) - r_N(R, P)) - (\tilde{e}_{\text{SP}}(R, P, r(R, P)) - r(R, P))\}^2}{2} \\
&\quad (89)
\end{aligned}$$

$$\begin{aligned}
&= \Lambda_{0,P}^*(\tilde{e}_{\text{SP}}(R, P, r(R, P)) - r(R, P)) + \epsilon_N \eta(R, P, r(R, P)) \\
&\quad + (\tilde{e}_{\text{SP}}(R, P, r_N(R, P)) - \tilde{e}_{\text{SP}}(R, P, r(R, P))) \eta(R, P, r(R, P)) \\
&\quad + \frac{\Lambda_{0,P}^{*''}(\bar{x}) \{(\tilde{e}_{\text{SP}}(R, P, r_N(R, P)) - r_N(R, P)) - (\tilde{e}_{\text{SP}}(R, P, r(R, P)) - r(R, P))\}^2}{2}, \\
&\quad (90)
\end{aligned}$$

where (89) follows by recalling the fact that $r_N(R, P) = r(R, P) - \epsilon_N$ and (90) follows from (88) by recalling (73).

Recalling the item (i) of the regularity lemma, i.e. Lemma 3.9, (90) implies that

$$\begin{aligned}
\Lambda_{0,P}^*(\tilde{e}_{\text{SP}}(R, P, r_N(R, P)) - r_N(R, P)) \tilde{e}_{\text{SP}}(R, P, r_N(R, P)) &= \tilde{e}_{\text{SP}}(R, P, r_N(R, P)) + \frac{\eta(R, P, r(R, P))}{1 - \eta(R, P, r(R, P))} \epsilon_N \\
&\quad + \frac{\Lambda_{0,P}^{*''}(\bar{x}) \epsilon_N^2}{2(1 - \eta(R, P, r(R, P)))} \left(1 + \frac{\tilde{e}_{\text{SP}}(R, P, r_N(R, P)) - \tilde{e}_{\text{SP}}(R, P, r(R, P))}{\epsilon_N}\right)^2, \\
&\quad (91)
\end{aligned}$$

for some $\bar{x} \in (\tilde{e}_{\text{SP}}(R, P, r(R, P)) - r(R, P), \tilde{e}_{\text{SP}}(R, P, r_N(R, P)) - r_N(R, P))$.

Note that, since $\tilde{e}_{\text{SP}}(R, P, \cdot) - (\cdot)$ is strictly decreasing and continuous, there exists a unique $\bar{r} \in (r(R, P) - \delta(R, \nu, W)/2, r(R, P))$ such that¹⁰ $\bar{x} = \tilde{e}_{\text{SP}}(R, P, \bar{r}) - \bar{r}$ and hence (recall (88) and (73))

$$\Lambda_{0,P}^{*''}(\bar{x}) = 1/\Lambda_{0,P}^{*''}(\eta(R, P, \bar{r})). \quad (92)$$

Moreover, the item (iii) of the regularity lemma, i.e. Lemma 3.9, implies that

$$\frac{\eta(R, P, r(R, P))}{1 - \eta(R, P, r(R, P))} = s^*(R, P, r(R, P)). \quad (93)$$

Plugging (92) and (93) into (91), we deduce that

$$\begin{aligned}
\Lambda_{0,P}^*(\tilde{e}_{\text{SP}}(R, P, r_N(R, P)) - r_N(R, P)) &= \tilde{e}_{\text{SP}}(R, P, r_N(R, P)) \\
&= \tilde{e}_{\text{SP}}(R, P, r(R, P)) + s^*(R, P, r(R, P)) \epsilon_N \\
&\quad + \frac{1 + s^*(R, P, r(R, P))}{2\Lambda_{0,P}^{*''}(\eta(R, P, \bar{r}))} \epsilon_N^2 \left(1 + \frac{\tilde{e}_{\text{SP}}(R, P, r_N(R, P)) - \tilde{e}_{\text{SP}}(R, P, r(R, P))}{\epsilon_N}\right)^2, \\
&\quad (94)
\end{aligned}$$

¹⁰Actually, $\bar{r} \in (r_N(R, P), r(R, P))$.

Moreover, using exactly the same arguments as above, but this time with a first-order Taylor series, we infer that

$$\tilde{e}_{\text{SP}}(R, P, r_N(R, P)) = \tilde{e}_{\text{SP}}(R, P, r(R, P)) + \epsilon_N \frac{\eta(R, P, \tilde{r})}{1 - \eta(R, P, \tilde{r})}, \quad (95)$$

for some $\tilde{r} \in (r_N(R, P), r(R, P))$.

On account of the convexity and the non-increasing property of $\tilde{e}_{\text{SP}}(R, P, \cdot)$, we have

$$\left| \frac{\partial \tilde{e}_{\text{SP}}(R, P, r')}{\partial r'} \right| \leq \frac{\tilde{e}_{\text{SP}}(R, P, 0)}{\delta(R, \nu, W)/2}, \quad (96)$$

for any $r_N(R, P) \leq r' \leq r(R, P)$.

By noting that $\tilde{e}_{\text{SP}}(R, P, 0) = D(W_{R,P}^- || W|P) = \Lambda'_{0,P}(1)$ and letting¹¹ $F := \max_{P \in \mathcal{P}_{R,\nu}(\mathcal{X})} \Lambda'_{0,P}(1) < \infty$, (96) further implies that

$$\frac{\eta(R, P, r')}{1 - \eta(R, P, r')} = s^*(R, P, r') = \left| \frac{\partial \tilde{e}_{\text{SP}}(R, P, r')}{\partial r'} \right| \leq \frac{F}{\delta(R, \nu, W)/2} =: \tilde{s} < \infty, \quad (97)$$

for any $r_N(R, P) \leq r' \leq r(R, P)$.

Plugging (81), (95) and (97) into (94) yields

$$\begin{aligned} \Lambda_{0,P}^*(\tilde{e}_{\text{SP}}(R, P, r_N(R, P)) - r_N(R, P)) &= \tilde{e}_{\text{SP}}(R, P, r_N(R, P)) \\ &\leq \tilde{e}_{\text{SP}}(R, P, r(R, P)) \\ &+ s^*(R, P, r(R, P))\epsilon_N \left[1 + \epsilon_N \frac{(1 + \tilde{s})^2 [1 + s^*(R, P, r(R, P))]}{2\underline{V}(\nu, W, R)s^*(R, P, r(R, P))} \right] \\ &= E_{\text{SP}}(R, P) + s^*(R, P, r(R, P))\epsilon_N \left[1 + \epsilon_N \frac{(1 + \tilde{s})^2 [1 + s^*(R, P, r(R, P))]}{2\underline{V}(\nu, W, R)s^*(R, P, r(R, P))} \right] \\ &\leq E_{\text{SP}}(R, P) + s^*(R, P, r(R, P))\epsilon_N \left[1 + \epsilon_N \frac{(1 + \tilde{s})^2}{2\underline{V}(\nu, W, R)} \left(1 + \frac{2\Upsilon(W, R, \nu)}{\nu} \right) \right], \end{aligned} \quad (98)$$

where (98) follows from the equality of the exponents theorem, i.e. Theorem 3.1 and the shifted exponent lemma, i.e. Lemma 3.4, and (99) follows from (93) and Lemma 3.10.

Consider $\zeta \in \mathbb{R}^+$ that is fixed in the definition of ϵ_N . Since \tilde{s} is bounded, $\underline{V}(\nu, W, R)$ and $\Upsilon(W, R, \nu)$ and the fact that $\nu > 0$, one can deduce that for all sufficiently large N ,

$$\epsilon_N \frac{(1 + \tilde{s})^2}{2\underline{V}(\nu, W, R)} \left(1 + \frac{2\Upsilon(W, R, \nu)}{\nu} \right) \leq \zeta,$$

and hence (99) reduces to the following, for all sufficiently large N ,

$$\Lambda_{0,P}^*(\tilde{e}_{\text{SP}}(R, P, r_N(R, P)) - r_N(R, P)) \leq E_{\text{SP}}(R, P) + s^*(R, P, r(R, P))\epsilon_N(1 + \zeta). \quad (100)$$

Next, we claim that

$$s^*(R, P, r(R, P)) = \rho_{R,P}^*. \quad (101)$$

¹¹Owing to the continuity lemma, i.e. Lemma 3.6, the maximum is well-defined and finite.

To prove this, we first claim that $\rho_{R,P}^*$ is a Lagrange multiplier of $e_{\text{SP}}(R, P)$. To see this, first note that

$$e_{\text{SP}}(R, P) = E_{\text{SP}}(R, P) \quad (102)$$

$$= K_{R,P}(\rho_{R,P}^*, Q_{R,P}^*) \quad (103)$$

$$= \max_{\rho \in \mathbb{R}_+} K_{R,P}(\rho, Q_{R,P}^*) \quad (104)$$

$$= \max_{\rho \in \mathbb{R}_+} \min_{V \in \mathcal{P}(\mathcal{Y}|\mathcal{X})} [\mathbf{D}(V||W|P) + \rho(\mathbf{D}(V||Q_{R,P}^*|P) - R)], \quad (105)$$

where (102) follows from the equality of the exponents theorem, i.e. Theorem 3.1, (103) follows from the saddle-point proposition, i.e. Proposition 3.1, and the uniqueness of the saddle-point proposition, i.e. Proposition 3.2, (104) follows by noting that $(\rho_{R,P}^*, Q_{R,P}^*)$ is the unique saddle-point of $K_{R,P}(\cdot, \cdot)$ and (105) follows by solving the convex minimization problem. Hence, (105) gives the Lagrangian dual of $e_{\text{SP}}(R, P)$.

Further, one can also check that

$$\max_{\rho \in \mathbb{R}_+} \min_{V \in \mathcal{P}(\mathcal{Y}|\mathcal{X})} [\mathbf{D}(V||W|P) + \rho(\mathbf{D}(V||Q_{R,P}^*|P) - R)] = \min_{V \in \mathcal{P}(\mathcal{Y}|\mathcal{X})} [\mathbf{D}(V||W|P) + \rho_{R,P}^*(\mathbf{D}(V||Q_{R,P}^*|P) - R)]. \quad (106)$$

(105) and (106) implies that $\rho_{R,P}^*$ is a Lagrange multiplier of $e_{\text{SP}}(R, P)$. Moreover, the sub-differential characterization of the Lagrange multipliers (e.g. [30, Theorem 29.1]) along with the differentiability of the shifted exponent lemma, i.e. Lemma 3.8, and Corollary 3.2, implies (101).

Plugging (101) into (100), we deduce that

$$\Lambda_{0,P}^*(\tilde{e}_{\text{SP}}(R, P, r_N(R, P)) - r_N(R, P)) \leq E_{\text{SP}}(R, P) + \rho_{R,P}^* \epsilon_N (1 + \zeta). \quad (107)$$

Define $\mathcal{P}_R^*(\mathcal{X}) := \{P \in \mathcal{P}(\mathcal{X}) : E_{\text{SP}}(R, P) = E_{\text{SP}}(R)\} \neq \emptyset$. Observe that \mathcal{P}_R^* is a compact set. Also, for any $P \in \mathcal{P}(\mathcal{X})$, $|P - \mathcal{P}_R^*| := \inf_{Q \in \mathcal{P}_R^*} \|Q - P\|_1$. For any $\theta \in \mathbb{R}^+$, $\mathcal{P}_\theta(\mathcal{X}) := \{P \in \mathcal{P}_{R,\nu}(\mathcal{X}) : |P - \mathcal{P}_R^*(\mathcal{X})| \geq \theta\}$.

Observe that (recall (1) and the differentiability of $E_{\text{SP}}(\cdot, P)$ proposition, i.e. Proposition 3.3)

$$\rho_R^* = \max_{P \in \mathcal{P}_R^*(\mathcal{X})} \rho_{R,P}^*, \quad (108)$$

where owing to the compactness of $\mathcal{P}_R^*(\mathcal{X})$ and the continuity of $\rho_{R,\cdot}^*$, the maximum is well-defined and finite.

Since $\mathcal{P}_{R,\nu}(\mathcal{X})$ is compact, $\rho_{R,\cdot}^*$ is uniformly continuous on this set, equivalently

$$\forall v \in \mathbb{R}^+, \exists a(v) \in \mathbb{R}^+, \text{ s.t. } \forall P, Q \in \mathcal{P}_{R,\nu}(\mathcal{X}), \|P - Q\|_1 < a(v) \Rightarrow |\rho_{R,P}^* - \rho_{R,Q}^*| < \zeta. \quad (109)$$

Consider $\zeta \in \mathbb{R}^+$ that is fixed in the definition of ϵ_N and let $a(\zeta) \in \mathbb{R}^+$ be chosen such that (109) holds.

If $P \in \mathcal{P}_{R,\nu}(\mathcal{X}) - \mathcal{P}_{a(\zeta)}(\mathcal{X})$, then (109) ensures that $\rho_{R,P}^* \leq \rho_R^* + \zeta$, which, in turn, implies that

$$\exp(-N\epsilon_N(1 + \zeta)\rho_{R,P}^*) \geq N^{-(1+\zeta)(\frac{1}{2}+\zeta)(\rho_R^*+\zeta)}. \quad (110)$$

Suppose $P \in \mathcal{P}_{a(\zeta)}(\mathcal{X})$. Since $E_{\text{SP}}(R) - \max_{P \in \text{cl}(\mathcal{P}_{a(\zeta)})} E_{\text{SP}}(R, P) \in \mathbb{R}^+$, one can check that for all sufficiently large N , uniformly over $\mathcal{P}_{a(\zeta)}(\mathcal{X})$, we have

$$\exp(-N[E_{\text{SP}}(R, P) + \epsilon_N(1 + \zeta)\rho_{R,P}^*]) \geq \frac{e^{-NE_{\text{SP}}(R)}}{N^{(1+\zeta)(\frac{1}{2}+\zeta)\rho_R^*}}. \quad (111)$$

Equations (87), (107), (110) and (111) imply (3), hence we conclude the proof of Theorem 2.1.

APPENDIX A
PROOF OF PROPOSITION 3.1

Lemma A.1: For any $R > R_\infty$ and $P \in \mathcal{P}(\mathcal{X})$,

$$\text{Esp}(R, P) = \max_{\rho \in \mathbb{R}_+} \min_{Q \in \mathcal{P}(\mathcal{Y})} \left\{ -\rho R - (1 + \rho) \Lambda_{Q,P} \left(\frac{\rho}{1 + \rho} \right) \right\}. \quad (112)$$

Proof: The proof is clear from basic optimization theoretic arguments, (e.g. [12, Exercise 2.5.23]), we just reproduce the steps for the sake of completeness.

$$\begin{aligned} \text{Esp}(R, P) &= \max_{\rho \in \mathbb{R}_+} \min_{V \in \mathcal{P}(\mathcal{Y}|\mathcal{X})} \{ \mathbf{D}(V||W|P) + \rho [\mathbf{I}(P; V) - R] \} \\ &= \max_{\rho \in \mathbb{R}_+} \min_{V \in \mathcal{P}(\mathcal{Y}|\mathcal{X})} \left\{ \mathbf{D}(V||W|P) + \rho \left[\min_{Q \in \mathcal{P}(\mathcal{Y})} \mathbf{D}(V||Q|P) - R \right] \right\} \\ &= \max_{\rho \in \mathbb{R}_+} \left\{ -\rho R + \min_{Q \in \mathcal{P}(\mathcal{Y})} \min_{V \in \mathcal{P}(\mathcal{Y}|\mathcal{X})} [\mathbf{D}(V||W|P) + \rho \mathbf{D}(V||Q|P)] \right\} \\ &= \max_{\rho \in \mathbb{R}_+} \min_{Q \in \mathcal{P}(\mathcal{Y})} \left\{ -\rho R - (1 + \rho) \Lambda_{Q,P} \left(\frac{\rho}{1 + \rho} \right) \right\}. \end{aligned} \quad (113)$$

■

Remark A.1: Recalling the definitions of $\mathcal{P}_{P,W}(\mathcal{Y})$ and $\tilde{\mathcal{P}}_{P,W}(\mathcal{Y})$ (cf. (12) and (13)), we note the following facts:

- (i) $\mathcal{P}_{P,W}(\mathcal{Y})$ and $\tilde{\mathcal{P}}_{P,W}(\mathcal{Y})$ are convex sets and $\tilde{\mathcal{P}}_{P,W}(\mathcal{Y}) \subset \mathcal{P}_{P,W}(\mathcal{Y})$.
- (ii) From the basic facts about convex sets (e.g. [29, Proposition 1.4.1 (c), Proposition 1.4.3 (b)]), $\text{ri}(\mathbb{R}_+) = \mathbb{R}^+$ and $\text{ri}(\mathcal{P}_{P,W}(\mathcal{Y})) = \text{ri}(\mathcal{P}(\mathcal{Y})) = \{Q \in \mathcal{P}(\mathcal{Y}) : Q(y) > 0, \forall y \in \mathcal{Y}\}$.
- (iii) For any $Q \in \mathcal{P}_{P,W}(\mathcal{Y})$, $\Lambda_{Q,P}(\lambda) \in \mathbb{R}$, for all $\lambda \in [0, 1)$.
- (iv) For any $Q \in \mathcal{P}(\mathcal{Y}) \setminus \mathcal{P}_{P,W}(\mathcal{Y})$, $\Lambda_{Q,P}(\lambda) = -\infty$, for all $\lambda \in (0, 1)$ and hence given any $R > R_\infty$, $P \in \mathcal{P}(\mathcal{X})$ and $Q \in \mathcal{P}(\mathcal{Y}) \setminus \mathcal{P}_{P,W}(\mathcal{Y})$, $K_{R,P}(\rho, Q) = \infty$ for all $\rho \in \mathbb{R}^+$.

Lemma A.2: Consider any $R > R_\infty$ and $P \in \mathcal{P}(\mathcal{X})$.

- (i) Given any $\rho \in \mathbb{R}_+$ (resp. $\rho \in \mathbb{R}^+$), $K_{R,P}(\rho, \cdot)$ is (resp. strictly) convex on $\mathcal{P}_{P,W}(\mathcal{Y})$ (resp. $\tilde{\mathcal{P}}_{P,W}(\mathcal{Y})$).
- (ii) Given any $Q \in \mathcal{P}_{P,W}(\mathcal{Y})$, $K_{R,P}(\cdot, Q)$ is concave on \mathbb{R}_+ .

Proof: Let $R > R_\infty$ and $P \in \mathcal{P}(\mathcal{X})$ be arbitrary.

- (i) Given any $x \in \mathcal{S}(P)$ and $\lambda \in [0, 1)$ define $f_{x,\lambda} : \mathcal{P}_{P,W}(\mathcal{Y}) \rightarrow \mathbb{R}^+$ such that

$$f_{x,\lambda}(Q) := \begin{cases} \sum_{y \in \mathcal{Y}} W(y|x)^{1-\lambda} Q(y)^\lambda, & \text{if } \lambda \in (0, 1), \\ 1, & \text{if } \lambda = 0, \end{cases}$$

for any $Q \in \mathcal{P}_{P,W}(\mathcal{Y})$. Let $Q_1, Q_2 \in \mathcal{P}_{P,W}(\mathcal{Y})$ and $\theta \in (0, 1)$ be arbitrary. For any $\lambda \in (0, 1)$, we have

$$\begin{aligned} f_{x,\lambda}(\theta Q_1 + (1 - \theta) Q_2) &= \sum_{y \in \mathcal{Y}} W(y|x)^{1-\lambda} [\theta Q_1(y) + (1 - \theta) Q_2(y)]^\lambda \\ &\geq \sum_{y \in \mathcal{Y}} W(y|x)^{1-\lambda} [\theta Q_1(y)^\lambda + (1 - \theta) Q_2(y)^\lambda] \end{aligned} \quad (114)$$

$$= \theta f_{x,\lambda}(Q_1) + (1 - \theta) f_{x,\lambda}(Q_2), \quad (115)$$

where (114) follows from the concavity of $(\cdot)^\lambda$ on \mathbb{R}_+ for any $\lambda \in (0, 1)$. Clearly, (115) is true for $\lambda = 0$.

Since $\log(\cdot)$ is strictly increasing and strictly concave on \mathbb{R}^+ , (115) implies that

$$\begin{aligned} \log(f_{x,\lambda}(\theta Q_1 + (1-\theta)Q_2)) &\geq \log(\theta f_{x,\lambda}(Q_1) + (1-\theta)f_{x,\lambda}(Q_2)) \\ &\geq \theta \log(f_{x,\lambda}(Q_1)) + (1-\theta) \log(f_{x,\lambda}(Q_2)). \end{aligned} \quad (116)$$

(116) implies that given any $\rho \in \mathbb{R}_+$, $\Lambda_{\cdot,P}\left(\frac{\rho}{1+\rho}\right)$ is concave on $\mathcal{P}_{P,W}(\mathcal{Y})$. By recalling the definition of $K_{R,P}$ (cf. (14)), this implies that $K_{R,P}(\rho, \cdot)$ is convex on $\mathcal{P}_{P,W}(\mathcal{Y})$.

Strict concavity follows by noting that for any $Q_1, Q_2 \in \tilde{\mathcal{P}}_{P,W}(\mathcal{Y})$ such that $Q_1 \neq Q_2$ and $\lambda \in (0, 1)$, the inequality in (114) is strict owing to the strict concavity of $(\cdot)^\lambda$ on \mathbb{R}^+ for any $\lambda \in (0, 1)$.

(ii) For any $\lambda \in (0, 1)$, $Q \in \mathcal{P}_{P,W}(\mathcal{Y})$ and $x \in \mathcal{S}(P)$ define

$$\forall y \in \mathcal{Y}, \tilde{W}_{\lambda,Q}(y|x) := \frac{W(y|x)^{1-\lambda} Q(y)^\lambda}{\sum_{\tilde{y} \in \mathcal{Y}} W(\tilde{y}|x)^{1-\lambda} Q(\tilde{y})^\lambda}. \quad (117)$$

Recalling the definition of $\mathcal{P}_{P,W}(\mathcal{Y})$, $\tilde{W}_{\lambda,Q}(\cdot|x)$ is a well-defined probability measure on \mathcal{Y} . It is easy to check that¹²

$$\Lambda'_{Q,P}(\lambda) = \sum_{x \in \mathcal{S}(P)} P(x) \mathbb{E}_{\tilde{W}_{\lambda,Q}(\cdot|x)} \left[\log \frac{Q(Y)}{W(Y|x)} \right], \quad (118)$$

$$\Lambda''_{Q,P}(\lambda) = \sum_{x \in \mathcal{S}(P)} P(x) \text{Var}_{\tilde{W}_{\lambda,Q}(\cdot|x)} \left[\log \frac{Q(Y)}{W(Y|x)} \right], \quad (119)$$

for any $Q \in \mathcal{P}_{P,W}(\mathcal{Y})$ and $\lambda \in (0, 1)$. Recalling the definition of $K_{R,P}$ (cf. (14)), (119) implies that

$$\frac{\partial^2 K_{R,P}(\rho, Q)}{\partial \rho^2} = -\frac{1}{(1+\rho)^3} \Lambda''_{Q,P} \left(\frac{\rho}{1+\rho} \right) \leq 0, \quad (120)$$

for any $Q \in \mathcal{P}_{P,W}(\mathcal{Y})$ and $\rho \in \mathbb{R}^+$.

Now, fix any $Q \in \mathcal{P}_{P,W}(\mathcal{Y})$. (120) implies that $-K_{R,P}(\cdot, Q)$ is convex on \mathbb{R}^+ , equivalently, the epigraph of $-K_{R,P}(\cdot, Q)$ with its domain restricted to \mathbb{R}^+ is a convex set. Furthermore,

$$\lim_{\rho \downarrow 0} -K_{R,P}(\rho, Q) \leq 0 = -K_{R,P}(0, Q).$$

Hence, after adding 0 into the domain of $K_{R,P}(\cdot, Q)$, its epigraph remains to be convex. ■

Definition A.1: Let $G \subset \mathbb{R}^n$ and $f : G \rightarrow \mathbb{R}$. (G, f) is “convex and closed in Fenchel’s sense” (cf. [31, pg. 151], [32, end of Section 2]) (resp. “concave and closed in Fenchel’s sense”) provided that:

- (i) G is convex.
- (ii) f is convex (resp. concave) and lower (resp. upper) semi-continuous.
- (iii) Any accumulation point of G that does not belong to G satisfies $\lim f(\cdot) = \infty$ (resp. $\lim f(\cdot) = -\infty$).

Lemma A.3: Let $R > R_\infty$ and $P \in \mathcal{P}(\mathcal{X})$ be arbitrary. For any $Q \in \text{ri}(\mathcal{P}_{P,W}(\mathcal{Y}))$ (resp. $\rho \in \text{ri}(\mathbb{R}_+)$), $(\mathbb{R}_+, K_{R,P}(\cdot, Q))$ (resp. $(\mathcal{P}_{P,W}(\mathcal{Y}), K_{R,P}(\rho, \cdot))$) is concave (resp. convex) and closed in Fenchel’s sense.

¹²For the sake of notational convenience $\Lambda'_{Q,P}(\lambda)$ (resp. $\Lambda''_{Q,P}(\lambda)$) denotes $\frac{\partial \Lambda_{Q,P}(\lambda)}{\partial \lambda}$ (resp. $\frac{\partial^2 \Lambda_{Q,P}(\lambda)}{\partial \lambda^2}$) in the sequel.

Proof: Fix any $R > R_\infty$ and $P \in \mathcal{P}(\mathcal{X})$.

First, fix an arbitrary $Q \in \tilde{P}_{P,W}(\mathcal{Y})$. Observe that $\Lambda_{Q,P}(\lambda) \in \mathbb{R}$ for all $\lambda \in (0, 1)$, which in turn implies that $\Lambda_{Q,P}(\lambda)$ is infinitely differentiable with respect to λ for all $\lambda \in (0, 1)$. Moreover, recalling the definition of $\tilde{P}_{P,W}(\mathcal{Y})$, it is easy to check that for any $Q \in \tilde{P}_{P,W}(\mathcal{Y})$, $\lim_{\lambda \downarrow 0} \Lambda_{Q,P}(\lambda) = 0 = \Lambda_{Q,P}(0)$. These two observations ensure the continuity (and *a fortiori* upper semi-continuity) of $K_{R,P}(\cdot, Q)$ on \mathbb{R}_+ . By noting (recall item (ii) of Remark A.1) $\text{ri}(\mathcal{P}_{P,W}(\mathcal{Y})) \subset \tilde{P}_{P,W}(\mathcal{Y})$, the fact that \mathbb{R}_+ is closed and convex and the concavity of $K_{R,P}(\cdot, Q)$ (cf. item (ii) of Lemma A.2) this suffices to conclude that $(\mathbb{R}_+, K_{R,P}(\cdot, Q))$ is concave and closed in Fenchel's sense.

Next, fix an arbitrary $\rho \in \text{ri}(\mathbb{R}_+) = \mathbb{R}^+$ (cf. the item (ii) of Remark A.1). Observe that any accumulation point of $\mathcal{P}_{P,W}(\mathcal{Y})$ which does not belong to $\mathcal{P}_{P,W}(\mathcal{Y})$, say Q_0 , satisfies $Q_0 \in \mathcal{P}(\mathcal{Y}) \setminus \mathcal{P}_{P,W}(\mathcal{Y})$, owing to the compactness of $\mathcal{P}(\mathcal{Y})$, and hence $K_{R,P}(\rho, Q_0) = \infty$. Further, item (i) of Remark A.1 and item (i) of Lemma A.2 ensures that in order to conclude that $K_{R,P}(\rho, \cdot)$ is convex and closed in Fenchel's sense, we only need to verify the lower semi-continuity. Implied by its convexity, $K_{R,P}(\rho, \cdot)$ is continuous on $\text{ri}(\mathcal{P}(\mathcal{Y}))$. Let $Q_0 \in \mathcal{P}_{P,W}(\mathcal{Y}) \setminus \text{ri}(\mathcal{P}(\mathcal{Y}))$ be arbitrary. Consider an arbitrary sequence $\{Q_k\}_{k \geq 1}$ such that $Q_k \in \mathcal{P}_{P,W}(\mathcal{Y})$ and $\lim_{k \rightarrow \infty} Q_k = Q_0$. Lastly, define $\lambda := \frac{\rho}{1+\rho} \in (0, 1)$. We have

$$\begin{aligned} \lim_{k \rightarrow \infty} \Lambda_{Q_k,P}(\lambda) &= \lim_{k \rightarrow \infty} \sum_{x \in \mathcal{S}(P)} P(x) \log \sum_{y \in \mathcal{Y}} W(y|x)^{1-\lambda} Q_k(y)^\lambda \\ &= \sum_{x \in \mathcal{S}(P)} P(x) \log \sum_{y \in \mathcal{Y}} W(y|x)^{1-\lambda} Q_0(y)^\lambda \\ &= \Lambda_{Q_0,P}(\lambda), \end{aligned} \tag{121}$$

where (121) follows from the continuity of $\log(\cdot)$ and $(\cdot)^\lambda$. ■

Now, we are ready to prove the existence of a saddle-point. To this end, fix arbitrary $R > R_\infty$ and $P \in \mathcal{P}(\mathcal{X})$ from now on.

We first establish

$$-\infty < \max_{\rho \in \mathbb{R}_+} \inf_{Q \in \mathcal{P}_{P,W}(\mathcal{Y})} K_{R,P}(\rho, Q) = \min_{Q \in \mathcal{P}_{P,W}(\mathcal{Y})} \sup_{\rho \in \mathbb{R}_+} K_{R,P}(\rho, Q) < \infty. \tag{122}$$

In order to prove (122), we use a minimax theorem of Rockafellar, [31, Theorem 8]. Lemma A.3 ensures that $(\mathbb{R}_+, \mathcal{P}_{P,W}(\mathcal{Y}), K_{R,P})$ is a “closed saddle-element” (cf. [31, pg. 151]) and the boundedness of $\mathcal{P}_{P,W}(\mathcal{Y})$ guarantees the fulfillment of condition (II) for the validity of the aforementioned theorem (cf. [31, pg. 172]). Therefore [31, eq. (7.2)] implies that

$$-\infty < \sup_{\rho \in \mathbb{R}_+} \inf_{Q \in \mathcal{P}_{P,W}(\mathcal{Y})} K_{R,P}(\rho, Q) = \min_{Q \in \mathcal{P}_{P,W}(\mathcal{Y})} \sup_{\rho \in \mathbb{R}_+} K_{R,P}(\rho, Q). \tag{123}$$

Next, we claim that

$$\forall \rho \in \mathbb{R}_+, \inf_{Q \in \mathcal{P}_{P,W}(\mathcal{Y})} K_{R,P}(\rho, Q) = \inf_{Q \in \mathcal{P}(\mathcal{Y})} K_{R,P}(\rho, Q). \tag{124}$$

Since $\Lambda_{Q,P}(0) = 0$, for all $q \in \mathcal{P}(\mathcal{Y})$, (124) is trivially true for $\rho = 0$. On the other hand, for any $\rho \in \mathbb{R}^+$, item (iv) of Remark A.1 implies that

$$\forall Q \in \mathcal{P}(\mathcal{Y}) \setminus \mathcal{P}_{P,W}(\mathcal{Y}), K_{R,P}(\rho, Q) = \infty,$$

which, in turn, implies (124). Equation (112) and (124) imply that

$$E_{\text{SP}}(R, P) = \max_{\rho \in \mathbb{R}_+} \min_{Q \in \mathcal{P}(\mathcal{Y})} K_{R,P}(\rho, Q) = \max_{\rho \in \mathbb{R}_+} \inf_{Q \in \mathcal{P}_{P,W}(\mathcal{Y})} K_{R,P}(\rho, Q) < \infty. \quad (125)$$

Equation (123) and (125) imply that

$$-\infty < \max_{\rho \in \mathbb{R}_+} \inf_{Q \in \mathcal{P}_{P,W}(\mathcal{Y})} K_{R,P}(\rho, Q) = \min_{Q \in \mathcal{P}_{P,W}(\mathcal{Y})} \sup_{\rho \in \mathbb{R}_+} K_{R,P}(\rho, Q) < \infty,$$

which is (122).

From [30, Lemma 36.2], (122) ensures the existence of a saddle-point on $\mathbb{R}_+ \times \mathcal{P}_{P,W}(\mathcal{Y})$ and (125) implies the saddle-value is $E_{\text{SP}}(R, P)$. Hence we conclude the proof of the first assertion of the proposition.

Next, we prove the second assertion.

Lemma A.4: Consider any $R > R_\infty$ and $P \in \mathcal{P}(\mathcal{X})$. If $0 \in S(R, P)|_{\mathbb{R}_+}$, then $E_{\text{SP}}(R, P) = 0$, equivalently, if $E_{\text{SP}}(R, P) > 0$, then $0 \notin S(R, P)|_{\mathbb{R}_+}$.

Proof: Consider any $R > R_\infty$ and $P \in \mathcal{P}(\mathcal{X})$. Assume $0 \in S(R, P)|_{\mathbb{R}_+}$. We clearly have $K_{R,P}(0, Q) = 0$, for all $Q \in \mathcal{P}_{P,W}(\mathcal{Y})$, which in turn implies that (recall the definition of the saddle-point) $K_{R,P}(0, \hat{Q}) = 0$ for any $\hat{Q} \in \mathcal{P}_{P,W}(\mathcal{Y})$ satisfying $(0, \hat{Q}) \in S(R, P)$. From the first assertion of Proposition 3.1, this implies the claim. ■

Recalling the definition of $\mathcal{P}_R(\mathcal{X})$ (cf. (11)), Lemma A.4 immediately implies the following result.

Corollary A.1: For any $C > R > R_\infty$ and $P \in \mathcal{P}_R(\mathcal{X})$, $S(R, P)|_{\mathbb{R}_+} \subset \mathbb{R}^+$.

Lemma A.5: For any $C > R > R_\infty$ and $P \in \mathcal{P}_R(\mathcal{X})$, $S(R, P)|_{\mathcal{P}_{P,W}(\mathcal{Y})} \subset \tilde{\mathcal{P}}_{P,W}(\mathcal{Y})$.

Proof: Fix any $C > R > R_\infty$ and $P \in \mathcal{P}_R(\mathcal{X})$. Let $\hat{\rho} \in S(R, P)|_{\mathbb{R}_+}$ be arbitrary. Note that owing to Corollary A.1, $\hat{\rho} \in \mathbb{R}^+$. Define $\lambda := \frac{\hat{\rho}}{1+\hat{\rho}} \in (0, 1)$ and recall that (cf. proof of Lemma A.2) $\Lambda_{\cdot,P}(\lambda)$ is concave on $\mathcal{P}_{P,W}(\mathcal{Y})$.

For any $\hat{Q} \in \mathcal{P}_{P,W}(\mathcal{Y})$ such that $(\hat{\rho}, \hat{Q}) \in S(R, P)$ we have

$$K_{R,P}(\hat{\rho}, \hat{Q}) = \min_{Q \in \mathcal{P}_{P,W}(\mathcal{Y})} K_{R,P}(\hat{\rho}, Q) = -\hat{\rho}R - (1 + \hat{\rho}) \max_{Q \in \mathcal{P}_{P,W}(\mathcal{Y})} \Lambda_{Q,P} \left(\frac{\hat{\rho}}{1 + \hat{\rho}} \right), \quad (126)$$

from the definition of the saddle-point.

Now, consider any $Q \in \mathcal{P}_{P,W}(\mathcal{Y})$ and for any $x \in \mathcal{S}(P)$, define $\Lambda_{Q,x}(\lambda) := \log \sum_{y \in \mathcal{Y}} W(y|x)^{1-\lambda} Q(y)^\lambda$. Note that we have 3 possibilities for the partial derivatives of $\Lambda_{Q,x}(\lambda)$ with respect to $Q(y)$:

1) If $y \in \mathcal{S}(W(\cdot|x)) \cap \mathcal{S}(Q)$, then

$$\frac{\partial \Lambda_{Q,x}(\lambda)}{\partial Q(y)} = \frac{\lambda W(y|x)^{1-\lambda} Q(y)^{\lambda-1}}{\sum_{\tilde{y} \in \mathcal{Y}} W(\tilde{y}|x)^{1-\lambda} Q(\tilde{y})^\lambda}, \quad (127)$$

which is continuous in $Q(y)$.

2) If $y \notin \mathcal{S}(W(\cdot|x))$, then (since any variation along this direction does not change the value of the function)

$$\frac{\partial \Lambda_{Q,x}(\lambda)}{\partial Q(y)} = 0, \quad (128)$$

which is continuous in $Q(y)$.

3) If $y \notin \mathcal{S}(Q)$ and $y \in \mathcal{S}(W(\cdot|x))$, then

$$\frac{\partial \Lambda_{Q,x}(\lambda)}{\partial Q(y)} = \infty. \quad (129)$$

Then, [9, Theorem 4.4.1] implies that¹³ a necessary and sufficient condition for any $Q \in \mathcal{P}_{P,W}(\mathcal{Y})$ to achieve the maximum in (126) is:

$$\frac{\partial \Lambda_{Q,P}(\lambda)}{\partial Q(y)} = \delta, \forall y \in \mathcal{S}(Q), \quad (130)$$

$$\frac{\partial \Lambda_{Q,P}(\lambda)}{\partial Q(y)} \leq \delta, \forall y \notin \mathcal{S}(Q), \quad (131)$$

for some $\delta \in \mathbb{R}$. Clearly, if $Q \notin \tilde{\mathcal{P}}_{P,W}(\mathcal{Y})$ then it cannot satisfy (130) and (131) (cf. (129)). Hence, any minimizer of (126) belongs to $\tilde{\mathcal{P}}_{P,W}(\mathcal{Y})$. ■

Corollary A.1 and Lemma A.5 imply the second assertion of the proposition.

APPENDIX B PROOF OF PROPOSITION 3.2

Lemma B.1: Consider any $C > R > R_\infty$ and $P \in \mathcal{P}_R(\mathcal{X})$. For any $\hat{\rho} \in S(R, P)|_{\mathbb{R}_+}$, there exists a unique $\hat{Q} \in \mathcal{P}_{P,W}(\mathcal{Y})$, such that $(\hat{\rho}, \hat{Q}) \in S(R, P)$.

Proof: Consider any $C > R > R_\infty$ and $P \in \mathcal{P}_R(\mathcal{X})$. Let $\hat{\rho} \in S(R, P)|_{\mathbb{R}_+}$ be arbitrary. Existence of a $\hat{Q} \in \mathcal{P}_{P,W}(\mathcal{Y})$, such that $(\hat{\rho}, \hat{Q}) \in S(R, P)$ is guaranteed by the item (i) of saddle-point proposition, i.e. Proposition 3.1, hence we prove the uniqueness.

To this end, note that owing to the item (ii) of saddle-point proposition, (Corollary A.1 to be precise), $\hat{\rho} \in \mathbb{R}^+$. Moreover, the same result (Lemma A.5 to be precise) also implies that any $\hat{Q} \in \mathcal{P}_{P,W}(\mathcal{Y})$, such that $(\hat{\rho}, \hat{Q}) \in S(R, P)$ satisfies $Q \in \tilde{\mathcal{P}}_{P,W}(\mathcal{Y})$ and attains the minimum in the following expression

$$\min_{Q \in \tilde{\mathcal{P}}_{P,W}(\mathcal{Y})} K_{R,P}(\hat{\rho}, Q), \quad (132)$$

as a direct consequence of the definition of the saddle-point. However, item (i) of Lemma A.2 implies that $K_{R,P}(\hat{\rho}, \cdot)$ is strictly convex on $\tilde{\mathcal{P}}_{P,W}(\mathcal{Y})$ and hence the minimizer of (132) is unique. ■

Lemma B.2: Consider any $C > R > R_\infty$ and $P \in \mathcal{P}_R(\mathcal{X})$. For any $\hat{Q} \in S(R, P)|_{\mathcal{P}_{P,W}(\mathcal{Y})}$,

$$\forall \rho \in \mathbb{R}^+, \frac{\partial^2 K_{R,P}(\rho, \hat{Q})}{\partial \rho^2} = -\frac{1}{(1+\rho)^3} \Lambda''_{\hat{Q},P} \left(\frac{\rho}{1+\rho} \right) < 0, \quad (133)$$

and there exists a unique $\hat{\rho} \in \mathbb{R}_+$, such that $(\hat{\rho}, \hat{Q}) \in S(R, P)$.

Proof: Consider any $C > R > R_\infty$ and $P \in \mathcal{P}_R(\mathcal{X})$. Let $\hat{Q} \in S(R, P)|_{\mathcal{P}_{P,W}(\mathcal{Y})}$ be arbitrary. The existence of a $\hat{\rho} \in \mathbb{R}_+$, such that $(\hat{\rho}, \hat{Q}) \in S(R, P)$ is guaranteed by the item (i) of saddle-point proposition, i.e. Proposition 3.1, hence we prove the uniqueness.

¹³Strictly speaking the statement of the aforementioned theorem requires the cost function of the maximization problem to be continuously differentiable (with possible infinite value on the boundary) on the whole probability simplex. However, it is easy to verify that the proof given by Gallager is also applicable to our case. Indeed, for sufficiency, the item (iv) of Remark A.1 ensures that the value of the cost function evaluated at any Q satisfying (130) and (131) is not smaller than its counterpart for any $Q \in \mathcal{P}(\mathcal{Y}) \setminus \mathcal{P}_{P,W}(\mathcal{Y})$. For necessity, again the item (iv) of Remark A.1 ensures that any optimizer cannot be in $\mathcal{P}(\mathcal{Y}) \setminus \mathcal{P}_{P,W}(\mathcal{Y})$.

To this end, note that on account of the item (ii) of saddle-point proposition, (Lemma A.5, in particular), $\hat{Q} \in \tilde{\mathcal{P}}_{P,W}(\mathcal{Y})$, and hence $\Lambda_{\hat{Q},P}(\lambda)$ is infinitely differentiable with respect to λ on $(0, 1)$.

We first claim that

$$\Lambda''_{\hat{Q},P}(\lambda) > 0, \forall \lambda \in (0, 1). \quad (134)$$

For contradiction, suppose there exists a $\lambda \in (0, 1)$ such that $\Lambda''_{\hat{Q},P}(\lambda) = 0$. Note that

$$\begin{aligned} \left[\exists \lambda \in (0, 1), \text{ s.t. } \Lambda''_{\hat{Q},P}(\lambda) = 0 \right] &\Leftrightarrow \left[\exists \lambda \in (0, 1), \text{ s.t. } \sum_{x \in \mathcal{S}(P)} P(x) \text{Var}_{\tilde{W}_{\lambda, \hat{Q}}(\cdot|x)} \left[\log \frac{\hat{Q}(Y)}{W(Y|x)} \right] = 0 \right] \\ &\Leftrightarrow \left[\exists \lambda \in (0, 1), \text{ s.t. } \forall x \in \mathcal{S}(P), \text{Var}_{\tilde{W}_{\lambda, \hat{Q}}(\cdot|x)} \left[\log \frac{\hat{Q}(Y)}{W(Y|x)} \right] = 0 \right] \\ &\Leftrightarrow \left[\exists \lambda \in (0, 1), \text{ s.t. } \forall x \in \mathcal{S}(P), \hat{Q}(y) = W(y|x) e^{\Lambda'_{\hat{Q},x}(\lambda)}, \forall y \in \mathcal{S}(W(\cdot|x)) \right], \end{aligned} \quad (135)$$

where $\Lambda'_{\hat{Q},x}(\lambda) := \mathbb{E}_{\tilde{W}_{\lambda, \hat{Q}}(\cdot|x)} \left[\log \frac{\hat{Q}(Y)}{W(Y|x)} \right]$ (cf. (118)) and (135) follows from (119).

By the contradiction assumption, the left side of (135) is true. Fix any such $\lambda \in (0, 1)$. Then, for any $\rho \in \mathbb{R}^+$, we have

$$\begin{aligned} \Lambda_{\hat{Q},P} \left(\frac{\rho}{1+\rho} \right) &= \sum_{x \in \mathcal{S}(P)} P(x) \log \sum_{y \in \mathcal{S}(W(\cdot|x))} W(y|x)^{1/(1+\rho)} \hat{Q}(y)^{\rho/(1+\rho)}, \\ &= \frac{\rho}{1+\rho} \sum_{x \in \mathcal{S}(P)} P(x) \Lambda'_{\hat{Q},x}(\lambda), \end{aligned} \quad (137)$$

where (137) follows from (136). We further have,

$$\mathbb{E}_{\text{SP}}(R, P) = \max_{\rho \in \mathbb{R}_+} K_{R,P}(\rho, \hat{Q}), \quad (138)$$

$$= \max \left\{ 0, \sup_{\rho \in \mathbb{R}^+} K_{R,P}(\rho, \hat{Q}) \right\}, \quad (139)$$

where (138) follows by recalling the definition of the saddle-point and the item (i) of saddle-point proposition, i.e. Proposition 3.1, and (139) follows by noting the fact that $K_{R,P}(0, Q) = 0$ for all $Q \in \mathcal{P}_{P,W}(\mathcal{Y})$.

Also, (137) implies that

$$\begin{aligned} \sup_{\rho \in \mathbb{R}^+} K_{R,P}(\rho, \hat{Q}) &= \sup_{\rho \in \mathbb{R}^+} \left\{ -\rho R - \rho \sum_{x \in \mathcal{S}(P)} P(x) \Lambda'_{\hat{Q},x}(\lambda) \right\} \\ &= \sup_{\rho \in \mathbb{R}^+} -\rho \left\{ R + \Lambda'_{\hat{Q},P}(\lambda) \right\}, \end{aligned} \quad (140)$$

where (140) follows by recalling (118). Equations (139) and (140) clearly imply that either $\mathbb{E}_{\text{SP}}(R, P) = \infty$, which is impossible since $R > R_\infty$, or $\mathbb{E}_{\text{SP}}(R, P) = 0$, which is impossible since $P \in \mathcal{P}_R(\mathcal{X})$. Hence, (134) follows. A direct calculation reveals that (134) implies (133).

Next, recalling the definition of the saddle-point, we note that any $\hat{\rho} \in \mathbb{R}_+$ such that $(\hat{\rho}, \hat{Q}) \in S(R, P)$ satisfies

$$\begin{aligned} K_{R,P}(\hat{\rho}, \hat{Q}) &= \max_{\rho \in \mathbb{R}_+} K_{R,P}(\rho, \hat{Q}) \\ &= \max_{\rho \in \mathbb{R}^+} K_{R,P}(\rho, \hat{Q}), \end{aligned} \quad (141)$$

where (141) follows by recalling the assumption that $P \in \mathcal{P}_R(\mathcal{X})$. Equation (133) ensures that $K_{R,P}(\cdot, \hat{Q})$ is strictly concave on \mathbb{R}^+ and hence the maximizer of the right side of (141) is unique. ■

In order to conclude the proof, fix any $C > R > R_\infty$ and $P \in \mathcal{P}_R(\mathcal{X})$ and observe that (e.g. [33, Proposition VII.4.1.3]) $S(R, P) = S(R, P)|_{\mathbb{R}_+} \times S(R, P)|_{\mathcal{P}_{P,W}(\mathcal{Y})}$. Combining this fact with Lemmas B.1 and B.2 implies that $S(R, P)$ is a singleton, which was to be shown.

APPENDIX C PROOF OF PROPOSITION 3.3

First, we define the set of Lagrange multipliers of $E_{\text{SP}}(R, P)$ as follows: For any $R > R_\infty$ and $P \in \mathcal{P}(\mathcal{X})$,

$$\mathcal{L}(R, P) := \left\{ \hat{\rho} \in \mathbb{R}_+ : \hat{\rho} \text{ attains } \max_{\rho \in \mathbb{R}_+} \min_{V \in \mathcal{P}(\mathcal{Y}|\mathcal{X})} [\text{D}(V||W|P) + \rho(\text{I}(P; V) - R)] \right\}. \quad (142)$$

Lemma C.1: For any $R > R_\infty$ and $P \in \mathcal{P}(\mathcal{X})$, we have $\mathcal{L}(R, P) = S(R, P)|_{\mathbb{R}_+}$.

Proof: First of all, owing to the positivity of the relative entropy, it is easy to verify that

$$\text{I}(P; V) = \min_{Q \in \mathcal{P}(\mathcal{Y})} \text{D}(V||Q|P), \quad (143)$$

which, in turn, implies that (by solving the convex optimization problem)

$$\forall \rho \in \mathbb{R}_+, \min_{V \in \mathcal{P}(\mathcal{Y}|\mathcal{X})} \{ \text{D}(V||W|P) + \rho(\text{I}(P; V) - R) \} = \min_{Q \in \mathcal{P}(\mathcal{Y})} \left\{ -\rho R - (1 + \rho) \Lambda_{Q,P} \left(\frac{\rho}{1 + \rho} \right) \right\}. \quad (144)$$

Further, since for any $Q \in \mathcal{P}(\mathcal{Y})$, $\Lambda_{Q,P}(0) = 0$ and for any $\rho \in \mathbb{R}^+$, $\Lambda_{Q,P} \left(\frac{\rho}{1 + \rho} \right) = -\infty$, if $Q \notin \mathcal{P}_{P,W}(\mathcal{Y})$ (cf. item (iv) of Remark A.1), we have

$$\min_{Q \in \mathcal{P}(\mathcal{Y})} \left\{ -\rho R - (1 + \rho) \Lambda_{Q,P} \left(\frac{\rho}{1 + \rho} \right) \right\} = \inf_{Q \in \mathcal{P}_{P,W}(\mathcal{Y})} \left\{ -\rho R - (1 + \rho) \Lambda_{Q,P} \left(\frac{\rho}{1 + \rho} \right) \right\}. \quad (145)$$

Lastly, [30, Lemma 36.2] ensures that $\hat{\rho} \in S(R, P)|_{\mathbb{R}_+}$ if and only if $\hat{\rho}$ attains $\max_{\rho \in \mathbb{R}_+} \{ \inf_{Q \in \mathcal{P}_{P,W}(\mathcal{Y})} K_{R,P}(\rho, Q) \}$, which (owing to (144) and (145)) implies that $\mathcal{L}(R, P) = S(R, P)|_{\mathbb{R}_+}$. ■

Lemma C.2: For any $R > R_\infty$ and $P \in \mathcal{P}(\mathcal{X})$, we have $S(R, P)|_{\mathbb{R}_+} = -\partial E_{\text{SP}}(\cdot, P)(R)$, where $\partial E_{\text{SP}}(\cdot, P)(R)$ is the *subdifferential* of $E_{\text{SP}}(\cdot, P)$ at R (cf. [30, page 215]).

Proof: We note that (cf. [30, Theorem 29.1]¹⁴) $\mathcal{L}(R, P) = -\partial E_{\text{SP}}(\cdot, P)(R)$. The claim follows by recalling Lemma C.1. ■

Uniqueness of the saddle-point proposition, i.e. Proposition 3.2, and Lemma C.2 immediately imply that for any $C > R > R_\infty$ and $P \in \mathcal{P}_R(\mathcal{X})$,

$$S(R, P)|_{\mathbb{R}_+} = - \left. \frac{\partial E_{\text{SP}}(r, P)}{\partial r} \right|_{r=R}. \quad (146)$$

By recalling the definition of $\rho_{R,P}^*$ (e.g. (17)), (146) implies that

$$\rho_{R,P}^* = - \left. \frac{\partial E_{\text{SP}}(r, P)}{\partial r} \right|_{r=R},$$

which was to be shown.

¹⁴Strictly speaking, this result is stated for a finite dimensional Euclidean space. However, one can represent the stochastic matrices in $\mathbb{R}^{|\mathcal{X}||\mathcal{Y}|}$ and update each function accordingly and easily check this representation obeys the conditions of the aforementioned theorem. This reasoning applies to the similar situations in the sequel.

APPENDIX D
PROOF OF PROPOSITION 3.4

Let $C > R > R_\infty$ be arbitrary. Fix any $P_0 \in \mathcal{P}_R(\mathcal{X})$ and consider any $\{P_k\}_{k \geq 1}$ such that $P_k \in \mathcal{P}_R(\mathcal{X})$, $\forall k \in \mathbb{Z}^+$ and $\lim_{n \rightarrow \infty} P_k = P_0$.

We begin with showing the continuity of $\rho_{R,\cdot}^*$. Recalling (17) and the differentiability of $\text{E}_{\text{SP}}(\cdot, P)$ proposition, i.e. Proposition 3.3, we have

$$\forall k \in \mathbb{Z}_+, \rho_{R,P_k}^* = - \left. \frac{\partial \text{E}_{\text{SP}}(r, P_k)}{\partial r} \right|_{r=R}. \quad (147)$$

Further, continuity of $\text{E}_{\text{SP}}(\cdot, \cdot)$ on $(R_\infty, \infty) \times \mathcal{P}(\mathcal{X})$ (e.g. Lemma F.2) implies that

$$\lim_{k \rightarrow \infty} \text{E}_{\text{SP}}(R, P_k) = \text{E}_{\text{SP}}(R, P_0). \quad (148)$$

On account of (147), (148) and a continuity result of Hiriart-Urruty and Lemaréchal ([33, Corollary VI.6.2.8]) we conclude that

$$\lim_{k \rightarrow \infty} \rho_{R,P_k}^* = \rho_{R,P_0}^*,$$

which implies that $\rho_{R,\cdot}^*$ is continuous on $\mathcal{P}_R(\mathcal{X})$.

Next, we claim the continuity of $Q_{R,\cdot}^*$. Owing to the compactness of $\mathcal{P}(\mathcal{Y})$, there exists a subsequence $\{k_n\}_{n \geq 1}$ such that $\lim_{n \rightarrow \infty} Q_{R,P_{k_n}}^* = Q_0$ for some $Q_0 \in \mathcal{P}(\mathcal{Y})$. Consider such a subsequence.

Recalling the saddle-point proposition, i.e. Proposition 3.1, and the definitions of $\rho_{R,\cdot}^*$ and $Q_{R,\cdot}^*$ (e.g. (17) and (18)), we have

$$\forall n \in \mathbb{Z}^+, \text{E}_{\text{SP}}(R, P_{k_n}) = -R\rho_{R,P_{k_n}}^* - (1 + \rho_{R,P_{k_n}}^*)\Lambda_{Q_{R,P_{k_n}}^*, P_{k_n}} \left(\frac{\rho_{R,P_{k_n}}^*}{1 + \rho_{R,P_{k_n}}^*} \right). \quad (149)$$

Next, we define $f : \mathbb{R}_+ \times \mathbb{R}^+ \rightarrow \mathbb{R}$, such that $f(a, b) := a^b$ for any $(a, b) \in \mathbb{R}_+ \times \mathbb{R}^+$ and note that f is continuous on $\mathbb{R}_+ \times \mathbb{R}^+$. Using this, the continuity of $\rho_{R,\cdot}^*$ and $\log(\cdot)$, we deduce that

$$\lim_{n \rightarrow \infty} \Lambda_{Q_{R,P_{k_n}}^*, P_{k_n}} \left(\frac{\rho_{R,P_{k_n}}^*}{1 + \rho_{R,P_{k_n}}^*} \right) = \Lambda_{Q_0, P_0} \left(\frac{\rho_{R,P_0}^*}{1 + \rho_{R,P_0}^*} \right). \quad (150)$$

(149), (150) and the continuity of $\rho_{R,\cdot}^*$ implies that

$$\begin{aligned} -R\rho_{R,P_0}^* - (1 + \rho_{R,P_0}^*)\Lambda_{Q_0, P_0} \left(\frac{\rho_{R,P_0}^*}{1 + \rho_{R,P_0}^*} \right) &= \text{E}_{\text{SP}}(R, P_0) \\ &= \min_{Q \in \mathcal{P}_{P,W}(\mathcal{Y})} \left\{ -R\rho_{R,P_0}^* - (1 + \rho_{R,P_0}^*)\Lambda_{Q, P_0} \left(\frac{\rho_{R,P_0}^*}{1 + \rho_{R,P_0}^*} \right) \right\} \end{aligned} \quad (151)$$

$$= \min_{Q \in \mathcal{P}(\mathcal{Y})} \left\{ -R\rho_{R,P_0}^* - (1 + \rho_{R,P_0}^*)\Lambda_{Q, P_0} \left(\frac{\rho_{R,P_0}^*}{1 + \rho_{R,P_0}^*} \right) \right\}, \quad (152)$$

where (151) follows from recalling the definition of the saddle-point and (152) follows from item (iv) of Remark A.1.

The uniqueness of the saddle-point proposition, i.e. Proposition 3.2, the definition of $Q_{R,P}^*$ (e.g. (18)) and (152) imply that $Q_0 = Q_{R,P_0}^*$. Since $\{k_n\}_{n \geq 1}$ is arbitrary, we conclude that

$$\lim_{k \rightarrow \infty} Q_{R,P_k}^* = Q_{R,P_0}^*,$$

which implies that $Q_{R,\cdot}^*$ is continuous on $\mathcal{P}_R(\mathcal{X})$. Hence, we conclude the proof.

APPENDIX E PROOF OF THEOREM 3.1

Fix an arbitrary $C(W) > R > R_\infty$ and $P \in \mathcal{P}_R(\mathcal{X})$. Define $L(V, \rho) := D(V||W|P) + \rho(I(P; V) - R)$, for any $V \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$ and $\rho \in \mathbb{R}_+$. We have

$$E_{\text{SP}}(R, P) = \min_{V \in \mathcal{P}(\mathcal{Y}|\mathcal{X})} \sup_{\rho \in \mathbb{R}_+} L(V, \rho) = \max_{\rho \in \mathbb{R}_+} \min_{V \in \mathcal{P}(\mathcal{Y}|\mathcal{X})} L(V, \rho), \quad (153)$$

where the second equality follows from (113). (153) ensures that $L(\cdot, \cdot)$ has a saddle-point on $\mathcal{P}(\mathcal{Y}|\mathcal{X}) \times \mathbb{R}_+$. It is well-known that (e.g. [30, Corollary 28.3.1]) $\hat{V} \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$ is a minimizer of $E_{\text{SP}}(R, P)$ if and only if there exists some $\hat{\rho} \in \mathbb{R}_+$, such that $(\hat{V}, \hat{\rho})$ is a saddle-point of $L(\cdot, \cdot)$.

Recalling the definition of the saddle-point, the definition of $\rho_{R,P}^*$ (e.g. (17)), (142) and Lemma C.1, we conclude that an equivalent condition for $V_{R,P}^*$ to be an optimizer of $E_{\text{SP}}(R, P)$ is

$$V_{R,P}^* \in \arg \min_{V \in \mathcal{P}(\mathcal{Y}|\mathcal{X})} L(V, \rho_{R,P}^*). \quad (154)$$

Further,

$$E_{\text{SP}}(R, P) = \min_{V \in \mathcal{P}(\mathcal{Y}|\mathcal{X})} L(V, \rho_{R,P}^*) \quad (155)$$

$$= \min_{Q \in \mathcal{P}(\mathcal{Y})} \min_{V \in \mathcal{P}(\mathcal{Y}|\mathcal{X})} \{D(V||W|P) + \rho_{R,P}^*[D(V||Q|P) - R]\} \quad (156)$$

$$\leq \min_{V \in \mathcal{P}(\mathcal{Y}|\mathcal{X})} \{D(V||W|P) + \rho_{R,P}^*[D(V||Q_{R,P}^*|P) - R]\} \quad (157)$$

$$\leq K_{R,P}(\rho_{R,P}^*, Q_{R,P}^*) \quad (158)$$

$$= E_{\text{SP}}(R, P),$$

where (155) follows from (154), (156) follows from (143), (157) follows by plugging in $\tilde{W}^{\frac{\rho_{R,P}^*}{1+\rho_{R,P}^*}, Q_{R,P}^*}$ (cf. (117)) and (158) follows from the saddle-point proposition, i.e. Proposition 3.1 and the uniqueness of the saddle-point proposition, i.e. Proposition 3.2. Hence, we deduce that

$$\min_{V \in \mathcal{P}(\mathcal{Y}|\mathcal{X})} L(V, \rho_{R,P}^*) = \min_{V \in \mathcal{P}(\mathcal{Y}|\mathcal{X})} \{D(V||W|P) + \rho_{R,P}^*[D(V||Q_{R,P}^*|P) - R]\} = E_{\text{SP}}(R, P), \quad (159)$$

and $\tilde{W}^{\frac{\rho_{R,P}^*}{1+\rho_{R,P}^*}, Q_{R,P}^*}$ is an optimizer of $\min_{V \in \mathcal{P}(\mathcal{Y}|\mathcal{X})} \{D(V||W|P) + \rho_{R,P}^*[D(V||Q_{R,P}^*|P) - R]\}$. Moreover, since

$$L(V, \rho_{R,P}^*) \leq D(V||W|P) + \rho_{R,P}^*[D(V||Q_{R,P}^*|P) - R], \forall V \in \mathcal{P}(\mathcal{Y}|\mathcal{X}),$$

(159) further implies that $\tilde{W}^{\frac{\rho_{R,P}^*}{1+\rho_{R,P}^*}, Q_{R,P}^*} \in \arg \min_{V \in \mathcal{P}(\mathcal{Y}|\mathcal{X})} L(V, \rho_{R,P}^*)$, and hence $\tilde{W}^{\frac{\rho_{R,P}^*}{1+\rho_{R,P}^*}, Q_{R,P}^*}$ is a minimizer of $E_{\text{SP}}(R, P)$, owing to (154).

Next, we note that on account of (127), for any $Q \in \tilde{\mathcal{P}}_{P,W}(\mathcal{Y})$, we have

$$\frac{\partial \Lambda_{Q,P} \left(\frac{\rho_{R,P}^*}{1+\rho_{R,P}^*} \right)}{\partial Q(y)} = \frac{\rho_{R,P}^*}{1+\rho_{R,P}^*} \sum_{x \in \mathcal{S}(P)} P(x) \frac{W(y|x)^{1/(1+\rho_{R,P}^*)} Q(y)^{-1/(1+\rho_{R,P}^*)}}{\sum_{\tilde{y} \in \mathcal{Y}} W(\tilde{y}|x)^{1/(1+\rho_{R,P}^*)} Q(\tilde{y})^{\rho_{R,P}^*/(1+\rho_{R,P}^*)}}, \quad (160)$$

for all $y \in \mathcal{S}(Q)$. Moreover, (128) implies that for any $Q \in \tilde{\mathcal{P}}_{P,W}(\mathcal{Y})$,

$$\frac{\partial \Lambda_{Q,P} \left(\frac{\rho_{R,P}^*}{1+\rho_{R,P}^*} \right)}{\partial Q(y)} = 0, \forall y \notin \mathcal{S}(Q). \quad (161)$$

KKT conditions that $Q_{R,P}^*$ satisfies, i.e. (130) and (131), coupled with (160) and (161) (by choosing $\delta = \frac{\rho_{R,P}^*}{1+\rho_{R,P}^*}$ to ensure that $Q_{R,P}^*$ sums to 1) imply that

$$Q_{R,P}^*(y) = \sum_{x \in \mathcal{S}(P)} P(x) \frac{W(y|x)^{1/(1+\rho_{R,P}^*)} Q_{R,P}^*(y)^{\rho_{R,P}^*/(1+\rho_{R,P}^*)}}{\sum_{\tilde{y} \in \mathcal{Y}} W(\tilde{y}|x)^{1/(1+\rho_{R,P}^*)} Q_{R,P}^*(\tilde{y})^{\rho_{R,P}^*/(1+\rho_{R,P}^*)}}, \forall y \in \mathcal{Y}. \quad (162)$$

Clearly, (162) implies that

$$\sum_{x \in \mathcal{S}(P)} P(x) \tilde{W}_{\frac{\rho_{R,P}^*}{1+\rho_{R,P}^*}, Q_{R,P}^*}(y|x) = Q_{R,P}^*(y), \forall y \in \mathcal{Y},$$

which, in turn, implies that (since $\tilde{W}_{\frac{\rho_{R,P}^*}{1+\rho_{R,P}^*}, Q_{R,P}^*}$ is an optimizer of $E_{SP}(R, P)$)

$$I \left(P; \tilde{W}_{\frac{\rho_{R,P}^*}{1+\rho_{R,P}^*}, Q_{R,P}^*} \right) = D \left(\tilde{W}_{\frac{\rho_{R,P}^*}{1+\rho_{R,P}^*}, Q_{R,P}^*} \| Q_{R,P}^* | P \right) \leq R. \quad (163)$$

Next, we conclude the proof as follows. First,

$$\begin{aligned} e_{SP}(R, P) &= \inf_{V \in \mathcal{P}(\mathcal{Y}|\mathcal{X})} \sup_{\rho \in \mathbb{R}_+} \{ D(V||W|P) + \rho [D(V||Q_{R,P}^*|P) - R] \} \\ &\geq \inf_{V \in \mathcal{P}(\mathcal{Y}|\mathcal{X})} \{ D(V||W|P) + \rho_{R,P}^* [D(V||Q_{R,P}^*|P) - R] \} \\ &= E_{SP}(R, P), \end{aligned} \quad (164)$$

where (164) follows from (159).

On the other hand, (163) and the fact that $\tilde{W}_{\frac{\rho_{R,P}^*}{1+\rho_{R,P}^*}, Q_{R,P}^*}$ is a minimizer of $E_{SP}(R, P)$ ensure that

$$e_{SP}(R, P) \leq D \left(\tilde{W}_{\frac{\rho_{R,P}^*}{1+\rho_{R,P}^*}, Q_{R,P}^*} \| W | P \right) = E_{SP}(R, P). \quad (165)$$

Combining (164) and (165), we infer that

$$e_{SP}(R, P) = \min_{V \in \mathcal{P}(\mathcal{Y}|\mathcal{X}) : D(V||Q_{R,P}^*|P) \leq R} D(V||W|P) = E_{SP}(R, P),$$

which was to be shown.

APPENDIX F ANALYSIS OF THE CASE $P \in \mathcal{P}_{R,\nu}^c$

First, we define the following set: $\mathcal{P}_W(\mathcal{Y}|\mathcal{X}) := \{V \in \mathcal{P}(\mathcal{Y}|\mathcal{X}) : \forall x \in \mathcal{X}, V(\cdot|x) \ll W(\cdot|x)\}$. One can check the following via elementary calculations.

Lemma F.1: $\mathcal{P}_W(\mathcal{Y}|\mathcal{X})$ is convex and compact.

Next result will also be used in different parts of the paper.

Lemma F.2: $E_{SP}(\cdot, \cdot)$ is continuous on $(R_\infty, \infty) \times \mathcal{P}(\mathcal{X})$.

Proof: The proof follows similar lines to those of [12, Lemma 2.2.2], which proves continuity of the rate-distortion function.

First, note that given any $P \in \mathcal{P}(\mathcal{X})$, $E_{\text{SP}}(\cdot, P)$ is convex on (R_∞, ∞) . Fix an arbitrary $(R_0, P_0) \in (R_\infty, \infty) \times \mathcal{P}(\mathcal{X})$ and a sequence $\{(R_n, P_n)\}_{n \geq 1}$ such that $(R_n, P_n) \in (R_\infty, \infty) \times \mathcal{P}(\mathcal{X})$ and $\lim_{n \rightarrow \infty} (R_n, P_n) = (R_0, P_0)$.

Because of the convexity, $E_{\text{SP}}(\cdot, P_0)$ is continuous on (R_∞, ∞) . Hence, for any $\epsilon \in \mathbb{R}^+$ one can choose $V \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$ such that $I(P_0; V) < R_0$ and $D(V||W|P_0) < E_{\text{SP}}(R_0, P_0) + \epsilon$. Moreover, on account of continuity of $D(V||W|\cdot)$ and $I(\cdot; V)$, we have

$$D(V||W|P_n) < E_{\text{SP}}(R_0, P_0) + 2\epsilon, \quad I(P_n; V) \leq R_n,$$

for sufficiently large n , which, in turn, implies that

$$\limsup_{n \rightarrow \infty} E_{\text{SP}}(R_n, P_n) \leq E_{\text{SP}}(R_0, P_0). \quad (166)$$

Conversely, let $V_n \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$ be a minimizer of $E_{\text{SP}}(R_n, P_n)$ and w.l.o.g. suppose¹⁵ $V_n \in \mathcal{P}_W(\mathcal{Y}|\mathcal{X})$. Let $\{n_k\}_{k \geq 1}$ be a subsequence such that

$$\lim_{k \rightarrow \infty} E_{\text{SP}}(R_{n_k}, P_{n_k}) = \liminf_{n \rightarrow \infty} E_{\text{SP}}(R_n, P_n), \quad (167)$$

and

$$\lim_{k \rightarrow \infty} V_{n_k} = V, \quad (168)$$

for some $V \in \mathcal{P}_W(\mathcal{Y}|\mathcal{X})$. Note that existence of such a subsequence is ensured by the compactness of $\mathcal{P}_W(\mathcal{Y}|\mathcal{X})$ (cf. Lemma F.1). Equation (168) further implies that

$$\lim_{k \rightarrow \infty} I(P_{n_k}; V_{n_k}) = I(P_0; V) \leq R_0, \quad (169)$$

$$\lim_{k \rightarrow \infty} D(V_{n_k}||W|P_{n_k}) = D(V||W|P_0), \quad (170)$$

where (169) follows from the continuity of $I(\cdot; \cdot)$ and (170) follows from the continuity of $D(\cdot||W|\cdot)$ on $\mathcal{P}_W(\mathcal{Y}|\mathcal{X}) \times \mathcal{P}(\mathcal{X})$. Equations (167), (169) and (170) imply that

$$E_{\text{SP}}(R_0, P_0) \leq \liminf_{n \rightarrow \infty} E_{\text{SP}}(R_n, P_n). \quad (171)$$

Equations (166) and (171) imply that

$$\lim_{n \rightarrow \infty} E_{\text{SP}}(R_n, P_n) = E_{\text{SP}}(R_0, P_0).$$

■

Consider any $R_\infty < R < C$. For any $\nu \in \mathbb{R}^+$,

$$\mathcal{P}_{R,\nu}(\mathcal{X}) := \{P \in \mathcal{P}(\mathcal{X}) : E_{\text{SP}}(R, P) \geq \nu\}. \quad (172)$$

¹⁵To see why this does not yield a loss of generality, first note that since $E_{\text{SP}}(R_n, P_n) < \infty$, we necessarily have $V_n(\cdot|x) \ll W(\cdot|x)$, for all $x \in \mathcal{S}(P_n)$. On the other hand, $x \notin \mathcal{S}(P_n)$ does not affect neither the cost nor the constraint and hence the corresponding rows of the alternate channel, i.e. optimization variable of $E_{\text{SP}}(R_n, P_n)$, can be chosen arbitrarily without affecting optimality.

Let

$$\epsilon := (R - R_\infty)/2, \quad (173)$$

and fix an arbitrary $a \in (1, 2)$. Note that since $E_{\text{SP}}(\cdot)$ is convex, it is easy to see that it is Lipschitz continuous on $[R - \epsilon, R]$ (e.g. [30, Theorem 10.4]), i.e. there exists $L \in \mathbb{R}^+$, such that

$$\forall r_1, r_2 \in [R - \epsilon, R], \quad |E_{\text{SP}}(r_1) - E_{\text{SP}}(r_2)| \leq L|r_1 - r_2| \quad (174)$$

Next, we consider an arbitrary $\nu \in \mathbb{R}^+$ satisfying:

$$\nu \leq \min \left\{ (a-1), \frac{\epsilon}{2}, \frac{E_{\text{SP}}(R)(2-a)}{a(2L+1)} \right\}. \quad (175)$$

We claim that¹⁶

$$\max_{P \in \text{cl}(\mathcal{P}_{R,\nu}(\mathcal{X})^c)} E_{\text{SP}}(R - \nu, P) \leq \frac{E_{\text{SP}}(R)}{a}. \quad (176)$$

For contradiction, suppose

$$\max_{P \in \text{cl}(\mathcal{P}_{R,\nu}(\mathcal{X})^c)} E_{\text{SP}}(R - \nu, P) > \frac{E_{\text{SP}}(R)}{a}, \quad (177)$$

with a maximizer \tilde{P} . Since $E_{\text{SP}}(\cdot, \tilde{P})$ is convex and non-decreasing, (177) implies that

$$E_{\text{SP}}(R - 2\nu, \tilde{P}) > \frac{E_{\text{SP}}(R)}{a} + \nu \left(\frac{E_{\text{SP}}(R)}{a\nu} - 1 \right) = \frac{2E_{\text{SP}}(R)}{a} - \nu. \quad (178)$$

Further, owing to (175), we have

$$\frac{2E_{\text{SP}}(R)}{a} - \nu \geq E_{\text{SP}}(R) + 2L\nu. \quad (179)$$

Also, (174) and (175) imply that

$$E_{\text{SP}}(R - 2\nu) \leq E_{\text{SP}}(R) + 2L\nu. \quad (180)$$

Plugging (179) and (180) into (178) yields

$$E_{\text{SP}}(R - 2\nu, \tilde{P}) > E_{\text{SP}}(R - 2\nu),$$

which is a contradiction, by recalling the definition of $E_{\text{SP}}(\cdot)$, and hence (176) follows.

Let $P \in \text{cl}(\mathcal{P}_{R,\nu}(\mathcal{X})^c)$ be arbitrary. We have

$$(1 + \nu)E_{\text{SP}}(R - \nu, P) \leq \frac{(1 + \nu)E_{\text{SP}}(R)}{a} \quad (181)$$

$$\leq E_{\text{SP}}(R), \quad (182)$$

where (181) follows from (176) and (182) follows from (175).

Let (f, φ) be an (N, R) constant composition code with common composition $P \in \text{cl}(\mathcal{P}_{R,\nu}(\mathcal{X})^c)$. For all sufficiently large N , which only depends on $\nu, |\mathcal{X}|, |\mathcal{Y}|$, we have

$$e(f, \varphi) \geq \frac{1}{2} \exp(-N(1 + \nu)E_{\text{SP}}(R - \nu, P)) \quad (183)$$

$$\geq \frac{1}{2} \exp(-NE_{\text{SP}}(R)), \quad (184)$$

¹⁶Owing to Lemma F.2, the max is well-defined.

where (183) follows from the sphere packing lower bound for constant composition codes (cf. [12, Theorem 2.5.3]) and (184) follows from (182). Hence, we have the following lemma.

Lemma F.3: Fix $R_\infty < R < C$ and $\nu > 0$ satisfying (175). Then, for all sufficiently large N , which only depends on $\nu, |\mathcal{X}|$, and $|\mathcal{Y}|$, any (N, R) constant composition code with common composition $P \in \text{cl}(\mathcal{P}_{R,\nu}(\mathcal{X})^c)$ satisfies

$$e(f, \varphi) \geq \frac{1}{2} \exp(-N E_{\text{SP}}(R)). \quad (185)$$

APPENDIX G PROOF OF LEMMA 3.3

We begin with the proof of item (i). First, note that

$$\begin{aligned} D(V \| W_{R,P}^- | P) &= \sum_{x \in \mathcal{S}(P)} P(x) \sum_{y \in \mathcal{S}(V(\cdot|x))} V(y|x) \log \frac{V(y|x)}{W_{R,P}^-(y|x)} \\ &= \sum_{x \in \mathcal{S}(P)} P(x) \{ \log Q_{R,P}^* \{ \mathcal{S}(W(\cdot|x)) \} + D(V(\cdot|x) \| Q_{R,P}^*) \} \end{aligned} \quad (186)$$

$$= D(V \| Q_{R,P}^* | P) + \sum_{x \in \mathcal{S}(P)} P(x) \log Q_{R,P}^* \{ \mathcal{S}(W(\cdot|x)) \}, \quad (187)$$

where (186) follows from (33).

Similarly,

$$\begin{aligned} D(W_{R,P}^- \| Q_{R,P}^* | P) &= \sum_{x \in \mathcal{S}(P)} P(x) \sum_{y \in \mathcal{S}(W(\cdot|x))} W_{R,P}^-(y|x) \log \frac{W_{R,P}^-(y|x)}{Q_{R,P}^*(y)} \\ &= - \sum_{x \in \mathcal{S}(P)} P(x) \log Q_{R,P}^* \{ \mathcal{S}(W(\cdot|x)) \} \sum_{y \in \mathcal{S}(W(\cdot|x))} W_{R,P}^-(y|x) \end{aligned} \quad (188)$$

$$= - \sum_{x \in \mathcal{S}(P)} P(x) \log Q_{R,P}^* \{ \mathcal{S}(W(\cdot|x)) \}, \quad (189)$$

where (188) follows from the fact that $Q_{R,P}^* \in \tilde{\mathcal{P}}_{P,W}(\mathcal{Y})$ (cf. item (ii) of Proposition 3.1) and noting $W_{R,P}^-(\cdot|x) \equiv W(\cdot|x)$, for all $x \in \mathcal{X}$. Plugging (189) into (187) gives the item (i) of the lemma.

In order to prove the item (ii), observe that $(\rho_{R,P}^*, Q_{R,P}^*)$ is the unique saddle-point of $K_{R,P}(\cdot, \cdot)$. We have

$$\begin{aligned} K_{R,P}(\rho_{R,P}^*, Q_{R,P}^*) &= \max_{\rho \in \mathbb{R}_+} K_{R,P}(\rho, Q_{R,P}^*) \\ &= \max_{\rho \in \mathbb{R}^+} K_{R,P}(\rho, Q_{R,P}^*), \end{aligned} \quad (190)$$

where (190) follows by noting that $E_{\text{SP}}(R, P) = K_{R,P}(\rho_{R,P}^*, Q_{R,P}^*) > 0$ (cf. (158)) and $K_{R,P}(0, Q_{R,P}^*) = 0$.

Observe that $\rho_{R,P}^* \in \mathbb{R}^+$ is the unique maximizer of the right side of (190) and hence

$$\left. \frac{\partial K_{R,P}(\rho, Q_{R,P}^*)}{\partial \rho} \right|_{\rho=\rho_{R,P}^*} = -R - \Lambda_{Q_{R,P}^*, P} \left(\frac{\rho_{R,P}^*}{1 + \rho_{R,P}^*} \right) - \frac{1}{(1 + \rho_{R,P}^*)} \Lambda'_{Q_{R,P}^*, P} \left(\frac{\rho_{R,P}^*}{1 + \rho_{R,P}^*} \right) = 0. \quad (191)$$

Further,

$$\begin{aligned}
\lim_{\lambda \uparrow 1} \Lambda_{Q_{R,P}^*, P}(\lambda) &= \lim_{\lambda \uparrow 1} \sum_{x \in \mathcal{S}(P)} P(x) \log \sum_{y \in \mathcal{S}(W(\cdot|x))} W(y|x)^{1-\lambda} Q_{R,P}^*(y)^\lambda \\
&= \sum_{x \in \mathcal{S}(P)} P(x) \log \lim_{\lambda \uparrow 1} \sum_{y \in \mathcal{S}(W(\cdot|x))} W(y|x)^{1-\lambda} Q_{R,P}^*(y)^\lambda \\
&= \sum_{x \in \mathcal{S}(P)} P(x) \log \sum_{y \in \mathcal{S}(W(\cdot|x))} Q_{R,P}^*(y) \\
&= -D(W_{R,P}^- || Q_{R,P}^* | P),
\end{aligned} \tag{192}$$

where (192) follows from (189).

Moreover, recalling (25) and (26), for any $x \in \mathcal{S}(P)$

$$\lim_{\lambda \uparrow 1} \tilde{W}_{\lambda, Q_{R,P}^*}(y|x) = W_{R,P}^-(y|x), \tag{193}$$

for all $y \in \mathcal{Y}$. One can check that (e.g. (118))

$$\Lambda'_{Q_{R,P}^*, P}(\lambda) = \sum_{x \in \mathcal{S}(P)} P(x) \mathbb{E}_{\tilde{W}_{\lambda, Q_{R,P}^*}(\cdot|x)} \left[\log \frac{Q_{R,P}^*(Y)}{W(Y|x)} \right],$$

which, coupled with (193), implies that

$$\lim_{\lambda \uparrow 1} \Lambda'_{Q_{R,P}^*, P}(\lambda) = \sum_{x \in \mathcal{S}(P)} P(x) \sum_{y \in \mathcal{S}(W(\cdot|x))} W_{R,P}^-(y|x) \log \frac{Q_{R,P}^*(y)}{W(y|x)} \in \mathbb{R},$$

which, in turn, implies that

$$\lim_{\rho \rightarrow \infty} \frac{1}{(1+\rho)} \Lambda'_{Q_{R,P}^*, P} \left(\frac{\rho}{1+\rho} \right) = 0. \tag{194}$$

We have

$$0 > \lim_{\rho \rightarrow \infty} \frac{\partial K_{R,P}(\rho, Q_{R,P}^*)}{\partial \rho} \tag{195}$$

$$\begin{aligned}
&= \lim_{\rho \rightarrow \infty} -R - \Lambda_{Q_{R,P}^*, P} \left(\frac{\rho}{1+\rho} \right) - \frac{1}{(1+\rho)} \Lambda'_{Q_{R,P}^*, P} \left(\frac{\rho}{1+\rho} \right) \\
&= D(W_{R,P}^- || Q_{R,P}^* | P) - R,
\end{aligned} \tag{196}$$

where (195) follows from (191) and (133) and (196) follows from (192) and (194). Hence, we conclude that $R > D(W_{R,P}^- || Q_{R,P}^* | P)$.

APPENDIX H PROOF OF PROPOSITION 3.5

First, the property (i) above ensures that Λ_i is \mathcal{C}^∞ at η . Moreover, the property (ii) above implies that

$$\Lambda_n^*(q) = q\eta - \frac{1}{n} \sum_{i=1}^n \Lambda_i(\eta), \tag{197}$$

since $\frac{1}{n} \sum_{i=1}^n \Lambda_i(\delta)$ is convex.

Next, from (42), we have

$$\mathbb{E}_{\tilde{\lambda}_i} [Z_i] = \frac{1}{M_i(\eta)} \int z e^{\eta z} d\lambda_i(z).$$

Moreover, since Λ_i is \mathcal{C}^∞ at η , we also have

$$\Lambda'_i(\eta) = \frac{M'_i(\eta)}{M_i(\eta)} = \frac{1}{M_i(\eta)} \int z e^{\eta z} d\lambda_i(z).$$

And hence, we conclude that

$$\mathbb{E}_{\tilde{\lambda}_i}[Z_i] = \Lambda'_i(\eta). \quad (198)$$

Also, basic calculus reveals that

$$\Lambda''_i(\eta) = \frac{M''_i(\eta)}{M_i(\eta)} - \Lambda'_i(\eta)^2. \quad (199)$$

Moreover, since Λ_i is \mathcal{C}^∞ at η , we also have

$$M''_i(\eta) = \int z^2 e^{\eta z} d\lambda_i(z),$$

which, in turn, implies that (recall (42))

$$\mathbb{E}_{\tilde{\lambda}_i}[Z_i^2] = \frac{M''_i(\eta)}{M_i(\eta)}. \quad (200)$$

Plugging (198) and (200) into (199) yields

$$\text{Var}_{\tilde{\lambda}_i}[Z_i] = \Lambda''_i(\eta). \quad (201)$$

Furthermore, recalling (42), it is obvious that $\tilde{\lambda}_i \ll \lambda_i$. Moreover, since Z_i are real-valued and $e^{\eta z - \Lambda_i(\eta)} > 0$, for all $z \in \mathbb{R}$, we have

$$\frac{d\lambda_i}{d\tilde{\lambda}_i}(z) = e^{-\eta z + \Lambda_i(\eta)},$$

which, in turn, implies that $\lambda_i \ll \tilde{\lambda}_i$. Hence, we conclude that $\lambda_i \equiv \tilde{\lambda}_i$.

Next, we claim that

$$m_{2,n} > 0. \quad (202)$$

To see this, note that for any $i \in \{1, \dots, n\}$,

$$[\Lambda''_i(\eta) = 0] \iff [Z_i = \Lambda'_i(\eta) \quad \tilde{\lambda}_i - (\text{a.s.})] \quad (203)$$

$$\iff [Z_i = \Lambda'_i(\eta) \quad \lambda_i - (\text{a.s.})] \quad (204)$$

$$\implies [\text{Var}[Z_i] = 0], \quad (205)$$

where (203) follows from items (i) and (ii) of this remark, (204) follows since $\lambda_i \equiv \tilde{\lambda}_i$. From the assumption that $\sum_{i=1}^n \text{Var}[Z_i] > 0$ and (205), we conclude that $\sum_{i=1}^n \Lambda''_i(\eta) > 0$, which implies (202).

We continue as follows:

$$\begin{aligned} \mu_n([q, \infty)) &= \int_{\{\hat{S}_n \geq q\}} \lambda_1(dz_1) \dots \lambda_n(dz_n) \\ &= \int_{\{\hat{S}_n \geq q\}} e^{\sum_{i=1}^n [\Lambda_i(\eta) - \eta z_i]} \tilde{\lambda}_1(dz_1) \dots \tilde{\lambda}_n(dz_n) \end{aligned} \quad (206)$$

$$= e^{\sum_{i=1}^n \Lambda_i(\eta)} \mathbb{E}_{\tilde{\mu}_n} \left[\mathbb{1}_{\{\hat{S}_n \geq q\}} e^{-n\eta \hat{S}_n} \right] \quad (207)$$

$$= e^{-n\Lambda_n^*(q)} \mathbb{E}_{\tilde{\mu}_n} \left[\mathbb{1}_{\{\hat{S}_n \geq q\}} e^{-n[\eta \hat{S}_n - \eta q]} \right], \quad (208)$$

where (206) follows from (42), (207) follows by recalling the definition of $\tilde{\mu}_n$ and (208) follows from (197).

Note that (198) and (201) imply that

$$\mathbb{E}_{\tilde{\lambda}_i}[T_i] = 0, \quad \text{Var}_{\tilde{\lambda}_i}[T_i] = \Lambda_i''(\eta).$$

and note that

$$\hat{S}_n = \sqrt{m_{2,n}} \frac{W_n}{n} + q, \quad (209)$$

which, in turn, implies that

$$\{\hat{S}_n \geq q\} = \left\{ \sqrt{m_{2,n}} \frac{W_n}{n} \geq 0 \right\}. \quad (210)$$

Plugging (209) and (210) into (208) yields

$$\begin{aligned} \mu_n([q, \infty)) &= e^{-n\Lambda_n^*(q)} \mathbb{E}_{\tilde{\mu}_n} [\mathbb{1}_{\{W_n \geq 0\}} e^{-\eta \sqrt{m_{2,n}} W_n}] \\ &\geq e^{-n\Lambda_n^*(q)} \mathbb{E}_{\tilde{\mu}_n} [\mathbb{1}_{\{W_n \geq 0\}} e^{-\sqrt{m_{2,n}} W_n}] \end{aligned} \quad (211)$$

$$\begin{aligned} &= e^{-n\Lambda_n^*(q)} \int_0^\infty e^{-x \sqrt{m_{2,n}}} dF_n(x) \\ &= e^{-n\Lambda_n^*(q)} \int_0^\infty \sqrt{m_{2,n}} e^{-x \sqrt{m_{2,n}}} [F_n(x) - F_n(0)] dx \end{aligned} \quad (212)$$

$$= e^{-n\Lambda_n^*(q)} \int_0^\infty e^{-t} [F_n(t/\sqrt{m_{2,n}}) - F_n(0)] dt, \quad (213)$$

where F_n denotes the distribution function of W_n when Z_i are independent with the marginals $\tilde{\lambda}_i$, (211) follows from the fact that $\eta \leq 1$, (212) follows via integration by parts and (213) follows by letting $t := x\sqrt{m_{2,n}}$.

Note that since Λ_i is \mathcal{C}^∞ at η , $m_{3,n} < \infty$ and hence (recall (202)), $K_n(\eta) \in \mathbb{R}^+$.

Next, Berry-Esseen Theorem (cf. [34, Theorem III.1]. We use the particular instance of this theorem given by [34, eq. (III.15), pg. 43]) implies that

$$|F_n(x) - \Phi(x)| \leq c \frac{m_{3,n}}{m_{2,n}^{3/2}}, \quad \forall x \in \mathbb{R}, \quad (214)$$

where c is an absolute constant and can be chosen as $30/4$. Using (214), we deduce that

$$F_n(t/\sqrt{m_{2,n}}) \geq \Phi(t/\sqrt{m_{2,n}}) - \frac{cm_{3,n}}{m_{2,n}^{3/2}} \quad (215)$$

$$F_n(0) \leq \Phi(0) + \frac{cm_{3,n}}{m_{2,n}^{3/2}}. \quad (216)$$

Using (215) and (216) we get

$$\begin{aligned} F_n(t/\sqrt{m_{2,n}}) - F_n(0) &\geq \Phi(t/\sqrt{m_{2,n}}) - \Phi(0) - \frac{2cm_{3,n}}{m_{2,n}^{3/2}} \\ &\geq \phi(0) \frac{t}{\sqrt{m_{2,n}}} + \frac{\phi'(\bar{t})}{2} \left(\frac{t}{\sqrt{m_{2,n}}} \right)^2 - \frac{2cm_{3,n}}{m_{2,n}^{3/2}} \end{aligned} \quad (217)$$

$$\geq \frac{t}{\sqrt{2\pi m_{2,n}}} \left[1 - \frac{2cm_{3,n}\sqrt{2\pi}}{tm_{2,n}} \right] - \frac{1}{2\sqrt{2\pi}} \bar{t} e^{-\bar{t}^2/2} \frac{t^2}{m_{2,n}}, \quad (218)$$

where $0 \leq \bar{t} \leq t/\sqrt{m_{2,n}}$, (217) follows from Taylor's Theorem and (218) follows by noting $\phi'(x) = -\frac{x}{\sqrt{2\pi}}e^{-x^2/2}$.

Observe that $\mathbb{R}_+ \ni x \mapsto xe^{-x^2/2} \leq e^{-1/2} < 1$, which, in turn, implies that (recall (218))

$$F_n(t/\sqrt{m_{2,n}}) - F_n(0) \geq \frac{t}{\sqrt{2\pi m_{2,n}}} \left[1 - \frac{2cm_{3,n}\sqrt{2\pi}}{tm_{2,n}} \right] - \frac{t^2}{2\sqrt{2\pi m_{2,n}}}. \quad (219)$$

It is easy to check that

$$\int_{K_n(\eta)}^{\infty} te^{-t} dt = e^{-K_n(\eta)}(1 + K_n(\eta)), \quad (220)$$

$$\int_{K_n(\eta)}^{\infty} t^2 e^{-t} dt = e^{-K_n(\eta)}[1 + (1 + K_n(\eta))^2]. \quad (221)$$

Hence,

$$\begin{aligned} \int_0^{\infty} e^{-t} \left[F_n\left(\frac{t}{\sqrt{m_{2,n}}}\right) - F_n(0) \right] dt &\geq \int_{K_n(\eta)}^{\infty} e^{-t} \left[F_n\left(\frac{t}{\sqrt{m_{2,n}}}\right) - F_n(0) \right] dt \\ &\geq \frac{e^{-K_n(\eta)}}{\sqrt{2\pi m_{2,n}}} \left(1 - \frac{1 + (1 + K_n(\eta))^2}{2\sqrt{m_{2,n}}} \right), \end{aligned} \quad (222)$$

where (222) follows from (219), (220) and (221). Plugging (222) into (213) yields

$$\mu_n([q, \infty)) \geq \frac{e^{-n\Lambda_n^*(q)} e^{-K_n(\eta)}}{\sqrt{2\pi m_{2,n}}} \left(1 - \frac{1 + (1 + K_n(\eta))^2}{2\sqrt{m_{2,n}}} \right). \quad (223)$$

Clearly, if (43) holds, then (223) implies (44), which was to be shown.

APPENDIX I PROOF OF LEMMA 3.6

Let $(\lambda_0, P_0) \in (0, 1] \times \mathcal{P}_R(\mathcal{X})$ be arbitrary. Further, consider any $\{(\lambda_k, P_k)\}_{k \geq 1}$ such that $(\lambda_k, P_k) \in (0, 1] \times \mathcal{P}_R(\mathcal{X})$, for all $k \in \mathbb{Z}^+$ and $\lim_{k \rightarrow \infty} (\lambda_k, P_k) = (\lambda_0, P_0)$.

Note that for all sufficiently large $k \in \mathbb{Z}^+$, $\mathcal{S}(P_0) \subset \mathcal{S}(P_k)$. Consider such a $k \in \mathbb{Z}^+$. Recalling (49) and (50), we have

$$\Lambda'_{0,P_k}(\lambda_k) = \sum_{x \in \mathcal{S}(P_0)} P_k(x) \mathbb{E}_{\tilde{W}_{\lambda_k, P_k}(\cdot|x)} \left[\log \frac{W_{R, P_k}^-(Y|x)}{W(Y|x)} \right] + \sum_{x \in \mathcal{S}(P_0)^c} P_k(x) \mathbb{E}_{\tilde{W}_{\lambda_k, P_k}(\cdot|x)} \left[\log \frac{W_{R, P_k}^-(Y|x)}{W(Y|x)} \right]. \quad (224)$$

Using the continuity of the saddle-point proposition, i.e. Proposition 3.4, (25), (27) and the continuity of $\log(\cdot)$, it is easy to see that

$$\lim_{k \rightarrow \infty} P_k(x) \mathbb{E}_{\tilde{W}_{\lambda_k, P_k}(\cdot|x)} \left[\log \frac{W_{R, P_k}^-(Y|x)}{W(Y|x)} \right] = P_0(x) \mathbb{E}_{\tilde{W}_{\lambda_0, P_0}(\cdot|x)} \left[\log \frac{W_{R, P_0}^-(Y|x)}{W(Y|x)} \right], \quad \forall x \in \mathcal{S}(P_0),$$

which, in turn, implies that

$$\lim_{k \rightarrow \infty} \sum_{x \in \mathcal{S}(P_0)} P_k(x) \mathbb{E}_{\tilde{W}_{\lambda_k, P_k}(\cdot|x)} \left[\log \frac{W_{R, P_k}^-(Y|x)}{W(Y|x)} \right] = \sum_{x \in \mathcal{S}(P_0)} P_0(x) \mathbb{E}_{\tilde{W}_{\lambda_0, P_0}(\cdot|x)} \left[\log \frac{W_{R, P_0}^-(Y|x)}{W(Y|x)} \right]. \quad (225)$$

Next, we claim that

$$\lim_{k \rightarrow \infty} P_k(x) \mathbb{E}_{\tilde{W}_{\lambda_k, P_k}(\cdot|x)} \left[\log \frac{W_{R, P_k}^-(Y|x)}{W(Y|x)} \right] = 0, \quad (226)$$

for any $x \in \mathcal{S}(P_0)^c$. To see this, fix an arbitrary $x \in \mathcal{S}(P_0)^c$. If $x \in \mathcal{S}(P_k)$ for only finite number of k , then owing to (25), (226) is trivially true; hence suppose this is not the case. Let $\{k_n\}_{n \geq 1}$ be an arbitrary subsequence such that $x \in \mathcal{S}(P_{k_n})$, for all $n \in \mathbb{Z}^+$. Owing to the compactness of $\mathcal{P}(\mathcal{Y}|\mathcal{X})$ (switching to a subsubsequence if necessary) there exists $W_0(\cdot|x) \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$, such that

$$\lim_{n \rightarrow \infty} W_{R, P_{k_n}}^-(\cdot|x) = W_0(\cdot|x). \quad (227)$$

Since $W_{R, P_{k_n}}^-(\cdot|x) \ll W(\cdot|x)$ for all $n \in \mathbb{Z}^+$, it is easy to see that (cf. proof of Lemma F.1) $W_0(\cdot|x) \ll W(\cdot|x)$. This fact, along with the continuity of $\log(\cdot)$ and (227), implies that

$$\lim_{m \rightarrow \infty} \mathbb{E}_{\tilde{W}_{\lambda_{k_n}, P_{k_n}}(\cdot|x)} \left[\log \frac{W_{R, P_{k_n}}^-(Y|x)}{W(Y|x)} \right] = \mathbb{E}_{\tilde{W}_{\lambda_0, W_0(\cdot|x)}} \left[\log \frac{W_0(Y|x)}{W(Y|x)} \right] < \infty. \quad (228)$$

Noting $\lim_{m \rightarrow \infty} P_{k_{n_m}}(x) = P_0(x) = 0$ and the arbitrariness of the subsequence, (228) implies (226). Plugging (225) and (226) into (224) implies that

$$\lim_{k \rightarrow \infty} \Lambda'_{0, P_k}(\lambda_k) = \Lambda'_{0, P_0}(\lambda_0),$$

and hence we conclude $\Lambda'_{0, \cdot}(\cdot)$ is continuous on $(0, 1] \times \mathcal{P}_R(\mathcal{X})$.

By following exactly the same steps given above and noting the continuity of $(\cdot)^2$ (resp. $|\cdot|^3$), one can conclude the continuity of $\Lambda''_{0, \cdot}(\cdot)$ (resp. $m_{0,3}(\cdot, \cdot)$) on $(0, 1] \times \mathcal{P}_R(\mathcal{X})$.

Finally, the proof of the item (iv) follows from the similar arguments given in the proof of the item (i).

APPENDIX J PROOF OF LEMMA 3.9

Let $s^*(R, P, r) \in \mathbb{R}_+$ be as defined in (69). Since it is the unique maximizer of $\tilde{e}_{\text{SP}}(R, P, r)$, it should satisfy

$$r = \left. \frac{\partial e_o(s, P)}{\partial s} \right|_{s^*(R, P, r)}. \quad (229)$$

It is easy to verify that

$$\frac{\partial e_o(s, P)}{\partial s} = -\Lambda_{0, P} \left(\frac{s}{1+s} \right) - \frac{1}{1+s} \Lambda'_{0, P} \left(\frac{s}{1+s} \right), \quad (230)$$

Owing to (229) and (230), we have

$$r = -\Lambda_{0, P} \left(\frac{s^*(R, P, r)}{1+s^*(R, P, r)} \right) - \frac{1}{(1+s^*(R, P, r))} \Lambda'_{0, P} \left(\frac{s^*(R, P, r)}{1+s^*(R, P, r)} \right). \quad (231)$$

By noting (recall (65))

$$e_o(s^*(R, P, r), P) = -(1+s^*(R, P, r)) \Lambda_{0, P} \left(\frac{s^*(R, P, r)}{1+s^*(R, P, r)} \right),$$

Lemma 3.7, Corollary 3.1, (69) and (231) imply that

$$\tilde{e}_{\text{SP}}(R, P, r) = \frac{s^*(R, P, r)}{1+s^*(R, P, r)} \Lambda'_{0, P} \left(\frac{s^*(R, P, r)}{1+s^*(R, P, r)} \right) - \Lambda_{0, P} \left(\frac{s^*(R, P, r)}{1+s^*(R, P, r)} \right). \quad (232)$$

Due to (231) and (232), we get

$$\Lambda'_{0, P} \left(\frac{s^*(R, P, r)}{1+s^*(R, P, r)} \right) = \tilde{e}_{\text{SP}}(R, P, r) - r. \quad (233)$$

Using (71), (232) and (233), it is easy to see that (recall (197))

$$\Lambda_{0,P}^*(\tilde{e}_{\text{SP}}(R, P, r) - r) = \tilde{e}_{\text{SP}}(R, P, r),$$

which proves the item (i).

Item (ii) follows immediately from (53), (54), (71), (72) and the item (i).

In order to see the item (iii), first note that $\tilde{e}_{\text{SP}}(R, P, \cdot)$ is a non-increasing function. Further, it is clear that $\tilde{e}_{\text{SP}}(R, P, 0) = D(W_{R,P}^- || W | P)$ and $\tilde{e}_{\text{SP}}(R, P, D(W || W_{R,P}^- | P)) = 0$. These observations, along with (55), (56) and the positive variance lemma, i.e. Lemma 3.5, suffice to conclude the existence and uniqueness of $\eta(R, P, r) \in (0, 1)$ with the stated property. Finally, recalling (233), one can see that $\eta(R, P, r) = \frac{s^*(R, P, r)}{1 + s^*(R, P, r)}$, which completes the proof of the lemma.

REFERENCES

- [1] C. E. Shannon, "A Mathematical Theory of Communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.
- [2] P. Elias, "Coding For Two Noisy Channels," in *Information Theory, 3rd London Symp.*, 1955, pp. 61–76.
- [3] J. Wolfowitz, "The Coding of Messages Subject to Chance Errors," *Illinois J. of Math.*, vol. 1, no. 4, pp. 591–606, 1957.
- [4] V. Strassen, "Asymptotische Abschätzungen in Shannons Informationstheorie" *Trans. Third Prague Conf. Information Theory*, 1962, Czechoslovak Academy of Sciences, Prague, pp. 689–723.
- [5] R. M. Fano, *Transmission of Information, A Statistical Theory of Communications*. New York: Wiley, 1961.
- [6] R. G. Gallager, "A Simple Derivation of the Coding Theorem and Some Applications," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 3–18, Jan. 1965.
- [7] C. E. Shannon, R. G. Gallager and E. R. Berlekamp, "Lower Bounds to Error Probability for Coding on Discrete Memoryless Channels," *Inform. Contr.*, vol. 10, pp. 65–103, Jan. 1967.
- [8] E. A. Haroutunian, "Estimates of the Error Exponents for the Semi-Continuous Memoryless Channel," (in Russian) *Probl. Per. Inf.*, vol. 4, pp. 37–48, 1968.
- [9] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.
- [10] R. E. Blahut, *Principles and Practice of Information Theory*. Reading, MA: Addison-Wesley, 1987.
- [11] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
- [12] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic Press, 1981.
- [13] A. Valembois and M. Fossorier, "Sphere-Packing Bounds Revisited for Moderate Block Length," *IEEE Trans. Inform. Theory*, vol. 50, no. 12, pp. 2998–3014, Dec. 2004.
- [14] G. Wiechman and I. Sason, "An Improved Sphere-Packing Bound for Finite-Length Codes Over Symmetric Memoryless Channels," *IEEE Trans. Inform. Theory*, Vol. 54, no. 5, pp. 1962–1990, May 2008.
- [15] M. Hayashi, "Information Spectrum Approach to Second-Order Coding Rate in Channel Coding," *IEEE Trans. on Information Theory*, vol. IT 55, no. 11, pp. 4947–4966, November 2009.
- [16] Y. Polyanskiy, H. V. Poor and S. Verdú, "Channel Coding Rate in the Finite Blocklength Regime," *IEEE Trans. Inform. Theory*, vol. IT-56, no. 5, pp. 2307–2359, May 2010.
- [17] Y. Altuğ and A. B. Wagner, "Moderate Deviation Analysis of Channel Coding: Discrete Memoryless Case," in *Proc. 2010 IEEE Int. Symp. Inf. Theory*, June 2010, pp. 265–269.
- [18] Y. Polyanskiy and S. Verdú, "Channel Dispersion and Moderate Deviations Limits for Memoryless Channels," in *Proc. 48th Annu. Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Oct. 2010.
- [19] Y. Polyanskiy, H. V. Poor and S. Verdú, "Feedback in the Non-Asymptotic Regime," *IEEE Trans. on Information Theory*, vol. 57, no. 8, pp. 4903–4925, August 2011.
- [20] A. Ingber, R. Zamir and M. Feder, "Finite Dimensional Infinite Constellations," *arXiv:1103.0171*, September 2011.
- [21] Y. Altuğ and A. B. Wagner, "Refinement of the Sphere Packing Bound for Symmetric Channels," in *Proc. 49th Annu. Allerton Conf. Communication, Control, and Computing*, Monticello, IL, Oct. 2011.
- [22] Y. Altuğ and A. B. Wagner, "Refinement of the Random Coding Bound," in *Proc. Int. Zurich Sem. on Comm.*, Feb. 2012.
- [23] Y. Altuğ and A. B. Wagner, "Refinement of the Sphere-Packing Bound," in *Proc. 2012 IEEE Int. Symp. Inf. Theory*, July 2012.
- [24] Y. Altuğ and A. B. Wagner, "A Refinement of the Random Coding Bound," in *preparation*.
- [25] R. R. Bahadur and R. Ranga Rao, "On Deviations of the Sample Mean," *The Annals of Mathematical Statistics*, vol. 31, no. 4, pp. 1015–1027, Dec. 1960.
- [26] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications, 2nd edition*. New York: Springer-Verlag, 1998.
- [27] A. Ruszczyński, *Nonlinear Optimization*. Princeton, NJ: Princeton Univ. Press, 2006.
- [28] R. E. Blahut, "Hypothesis Testing and Information Theory," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 405–417, July 1974.
- [29] D. P. Bertsekas, A. Nedić and A. E. Ozdaglar, *Convex Analysis and Optimization*. Belmont, MA: Athena Scientific, 2003.
- [30] R. T. Rockafellar, *Convex Analysis*. Princeton, NJ: Princeton Univ. Press, 1970.
- [31] R. T. Rockafellar, "Minimax Theorems and Conjugate Saddle-Functions," *Math. Scand.*, vol. 14, pp. 151–173, 1964.
- [32] W. Fenchel, "On Conjugate Convex Functions," *Canad. J. Math.*, vol. 1, pp. 73–77, 1949.

- [33] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms I: Fundamentals*. Berlin: Springer-Verlag, 1993.
- [34] C.-G. Esseen, “Fourier analysis of distribution functions. A mathematical study of the Laplace–Gaussian law,” *Acta Math.*, vol. 77, pp. 1–125, 1945.